Data analysis in high-dimensional sparse spaces

Large p, small n problems

Clemmensen, Line Katrine Harder

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Data analysis in high-dimensional sparse spaces

Large p, small n problems

Line Clemmensen

Kongens Lyngby 2009
IMM-PHD-2009-228

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

The present thesis considers data analysis of problems with many features in relation to the number of observations (large p, small n problems). The theoretical considerations for such problems are outlined including the curses and blessings of dimensionality, and the importance of dimension reduction. In this context the trade off between a rich solution which answers the questions at hand and a simple solution which generalizes to unseen data is described. For all of the given data examples labelled output exists and the analyses are therefore limited to supervised settings.

Three novel classification techniques for high-dimensional problems are presented: Sparse discriminant analysis, sparse mixture discriminant analysis and orthogonality constrained support vector machines. The first two introduces sparseness to the well known linear and mixture discriminant analysis and thereby provide low-dimensional projections of data with few non-zero loadings which give improvements in classification. The latter adds *a priori* information of pairing between observations to the support vector machine and thereby give solutions with less variation and slight improvements in classification.

The classification methods are applied to classifications of fish species, ear canal impressions used in the hearing aid industry, microbiological fungi species, and various cancerous tissues and healthy tissues.

In addition, novel applications of sparse regressions (also called the elastic net) to the medical, concrete, and food industries via multi-spectral images for objective and automated systems are presented.

Resumé

Denne afhandling omhandler data analyse af problemer med mange variable og relativt få observationer (store p, små n problemer). De teoretiske overvejelser der er i forbindelse med sådanne problemer er beskrevet og disse inkluderer dimensionalitetens forbandelser og velsignelser og vigtigheden af at reducere dimensionerne. I den forbindelse beskrives afvejningen mellem en rig løsning som kan besvare de relevante spørgsmål og en simpel løsning som generaliserer til endnu usete observationer. I alle de givne eksempler findes der annoterede output variable og analyserne i afhandlingen er derfor begrænsede til superviserede analyser.

Tre nye klassifikationsteknikker for høj-dimensionale problemer præsenteres: Sparse discriminant analysis (sparsom diskriminant analyse), sparse mixture discriminant analysis (sparsom mikstur diskriminant analyse) og orthogonality constrained support vector machines (ortogonalitets begrænset support vektor maskiner). De første to metoder introducerer sparsomhed til de velkendte metoder lineær diskriminant analyse og mikstur diskriminant analyse og giver derved lav-dimensionale projektioner af data med få vægte forskellige fra nul hvilket medfører forbedringer i klassifikationerne. Den sidstnævnte metode tilfører *a priori* viden om parring af observationer til support vektor maskinen og dermed får løsningen en lavere varians og en smule bedre klassifikation.

Klassifikationsmetoderne er anvendt til at klassificere fiske arter, aftryk af øregange brugt i høreapparat industrien, mikrobiologiske svampe arter, og forskellige kræftangrebne vævsprøver og raske vævsprøver.

Ydermere præsenteres nye anvendelser af sparsom regression (også kaldet det elastiske net) til medicinal, beton og fødevare industrierne vha. multispektrale

billeder som bruges til at opnå objektive og automatiserede systemer.

Preface

This thesis was prepared at Informatics and Mathematical Modelling, the Technical University of Denmark in the Computer Graphics and Image Analysis group in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with statistical learning for problems with many variables in relation to the number of observations. The main focus is on extensions of supervised classification methods using regularization, but also regression problems and other dimension reduction techniques than regularization are considered.

The thesis consists of an introduction to the field of research, the basic methods utilized for the main focus in the thesis and a collection of seven research papers written during the period 2006–2009, and elsewhere published, and one unpublished report.

Lyngby, December 2009

Line H. Clemmensen

Papers included in the thesis

[A] Morten Mørup and Line Clemmensen. Multiplicative updates for the LASSO. In: *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2007. p. 33-38.

[B] Line Harder Clemmensen, Michael Edberg Hansen and Bjarne Kjær Ersbøll. A comparison of dimension reduction methods with applications to multispectral images of sand used in concrete. *Machine Vision* and *Applications, 2009. online first version (11 pages)*

[C] Line Clemmensen, Trevor Hastie and Bjarne Ersbøll. Sparse discriminant analysis. Technometrics. *resubmitted*

[D] Line Clemmensen and Sune Darkner. Classification of paired ear canal impressions in high dimensions - Data driven constraints for the Support Vector Machine Medical Image Analysis. *resubmitted*

[E] Line Harder Clemmensen and Bjarne Kjær Ersbøll. Multispectral recordings and analysis of psoriasis lesions. In: Proceedings of Workshop on Biophotonics Imaging for Diagnostics and Treatment, MICCAI, *2006. p. 15-18. IMM-Technical report-2006-17, Technical University of Denmark, DTU.*

[F] Line Harder Clemmensen and David Delgado Gomez and Bjarne Kjær Ersbøll. Individual discriminative face recognition models based on subsets of features. In: Proceedings of SCIA 2007 LNCS 4522*, p. 61-71, Ed. 2, Springer-Verlag, 2007.*

[G] Bjørn Dissing, Line Clemmensen, Hanne Løje, Bjarne Ersbøll and Jens Adler-Nissen. Temporal reflectance changes in vegetables, In: Proceed-

ings of IEEE workshop on Color and Reflectance in Imaging and computer Vision, CRICV, *2009*.

Other papers by the author

- Aloke Phatak, Harri Kiiveri, Line Harder Clemmensen, and William J. Wilson, NetRaVE: Constructing dependency networks using sparse linear regression, *Bioinformatics*, submitted, 2009.
- Line Harder Clemmensen, Michael Edberg Hansen, Bjarne Kjær Ersbøll, and Jens Christian Frisvad, A method for comparison of growth media in objective identification of Penicillium based on multi-spectral imaging, *Journal of Microbiological Methods*, 69: 249-255, 2007.
- David Delgado Gomez, Line Harder Clemmensen, Bjarne Kjær Ersbøll, and Jens Michael Carstensen, Precise acquisition and unsupervised segmentation of multi-spectral images, *Computer Vision and Image Understanding*, 106(2-3): 183-193, 2007.

Acknowledgements

I would like to thank all my fellow Ph.D. students and the staff in the Image Analysis and Computer Graphics Group at Informatics and Mathematical Modelling, Technical University of Denmark for good collaborations, a good atmosphere, and warm friendships throughout my Ph.D. studies.

I also thank my collaboration partners Dr. Michael Edberg Hansen at Pasteur Institute in Korea, and Prof. Jens Christian Frisvad at DTU Biosys for good discussions and good collaboration.

And to the CMIS group at CSIRO, in particular Prof. Harri Kiiveri, I send warm thanks for receiving me with open arms and hosting me for four months.

Likewise I would like to thank Prof. Trevor Hastie at The Statistics Department, Stanford University for good collaboration and for hosting me there for three days.

And of course warm thanks to my supervisor Bjarne Kjær Ersbøll for his great support and for guiding me through academia.

I thank my husband and my daughter for their love and all the joys they have and still are giving me, and for their support. *Meus amores, vocês estão sempre comigo no meu coraçao.*

Contents

Chapter 1

Introduction

In the last decade computer processors and memory have developed rapidly. It has become standard to take thousands of measurements such as metabolites, gene expressions, spectral wavelengths, or spatial image resolution and store them in large data bases. At the same time samples such as blood, fish, microbiological fungi, faces, or patients are not always accessible in the same numbers. In the present thesis we look at the interpretation of each such data set which has a large number of measurements also called features, (independent) variables, or inputs; depending on the literature, and a relatively small number of samples also called observations. The samples are in supervised settings related to a measurement which it is desirable to predict from a model based on the inputs. Such a measurement is called the dependent variable, the response, or the output; depending on the literature. The output can be a continuous measurement, or a categorical measurement (also called a class descriptor or class label).

1.1 Motivation

The research in this thesis is driven by learning from data. In the following subsections, the data problems considered throughout the thesis are described. In one way or another all the examples include more measurements than samples.

This is the common factor for the data examples and thus also for the research in the present thesis.

1.1.1 Example 1 - Moisture content in sand

The sand samples these data were based on came from the construction industry. The sand is used to make concrete, and it is of interest to know the moisture content of the sand in order to predict the right mixing proportions of the ingredients for the concrete. A construction line for such a process is illustrated in figure 1.1.

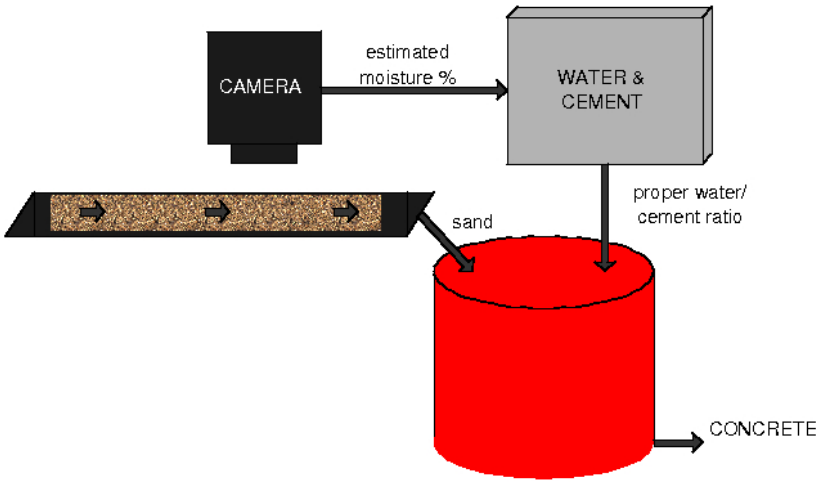
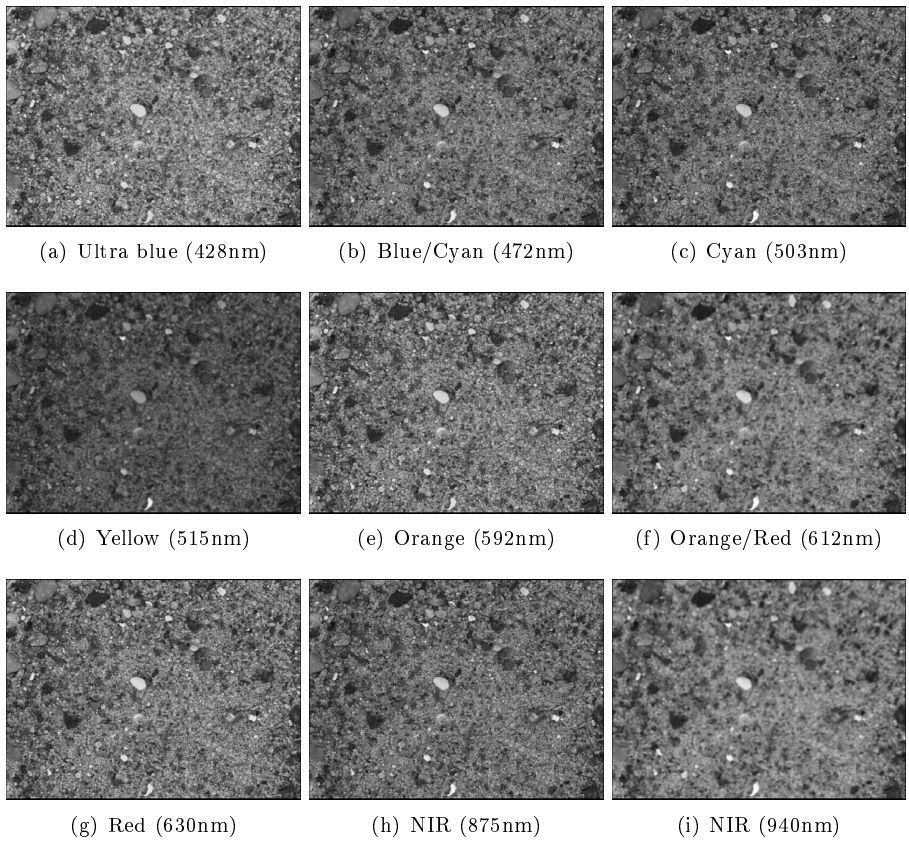


Figure 1.1: Diagram of the construction line in concrete mixing with digital estimation of the moisture content in sand.

The output is a measurement of the moisture content, and the input is multi-spectral images of the sand samples. In total 185 sand samples were digitized, and each multi-spectral image consists of $1035 \times 1380 \times 9$ spectral reflectance values. A standard RGB color image has three spectral bands: Red (650 nm), green (510 nm), and blue (475 nm), whereas a multi-spectral image consist of several spectral bands at various wavelengths of light . In this example there were 9 spectral bands; 7 in the visual range and 2 in the near infrared (NIR) range (see figure 1.2).

This data set is used in chapter 5 (paper B).



(a) Ultra blue (428nm) (b) Blue/Cyan (472nm) (c) Cyan (503nm)

(d) Yellow (515nm) (e) Orange (592nm) (f) Orange/Red (612nm)

(g) Red (630nm) (h) NIR (875nm) (i) NIR (940nm)

Figure 1.2: Illustration of a multi-spectral image of a sand sample. Each subfigure shows one of the spectral bands.

1.1.2 Example 2 - Ageing of vegetables

This data set consists of 12 multi-spectral images of celeriac and 12 of carrot samples. In each image there were around 30 pieces of preprocessed celeriac or carrot. The vegetables were digitalized over time, for 6 different days, and it is of interest to inspect the changes in texture and color over time. The texture and color changes are subsidiary variables for a quality measurement, or an assessment performed by a specialist. We used time as the output variable in this example. The multi-spectral images acquired for this data have 19 spectral bands, and thus each image contains $960 \times 1280 \times 19$ spectral reflectance values.

1.1.3 Example 3 - Recognition of faces

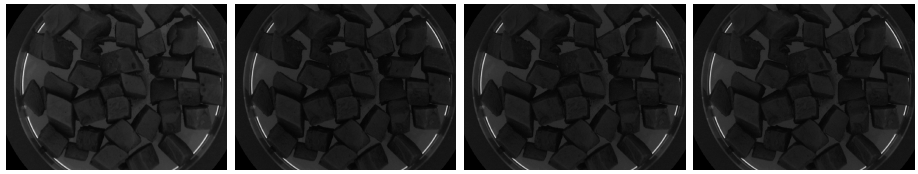
This data set is a subset of the XM2VTS (`http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`) database and consists of color images of the faces of 50 persons. For each person 8 images were regarded; resulting in a total of 400 RGB images. Each image consists of $\times 3$ color reflectance values and an example can be seen in figure 1.4. The task is to classify the face images into groups of the 50 persons.

1.1.4 Example 4 - Severity scoring of psoriasis lesions

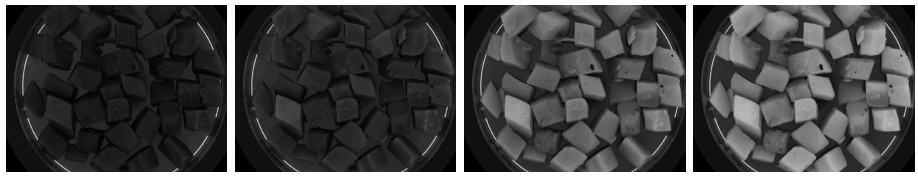
In this example there are 26 multi-spectral images of psoriasis lesions from 4 patients. A specialist evaluated the severity of the lesions according to the psoriasis area and severity index (PASI; Fredriksson and Petersson (1978)) on a scale from 0 to 4. The severity evaluation is the output of interest (also called a biomarker), and the multi-spectral images are the inputs. The images have a resolution of $1035 \times 1380 \times 9$ spectral reflectance values. The objective is to have an evaluation of the severity of psoriasis lesions which is reproducible from subject to subject and over time. These evaluations are of great importance to the choice of treatment.

1.1.5 Example 5 - Fish species

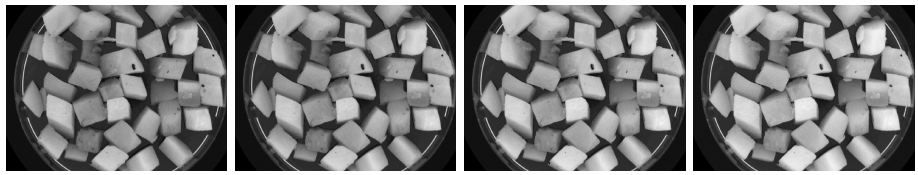
In the fishery industry the classification of fish species is in general performed manually. In Danish waters a lot of fish belonging to the Cod family are caught. In particular, the species: Cod, Haddock, and Whiting are difficult to tell apart



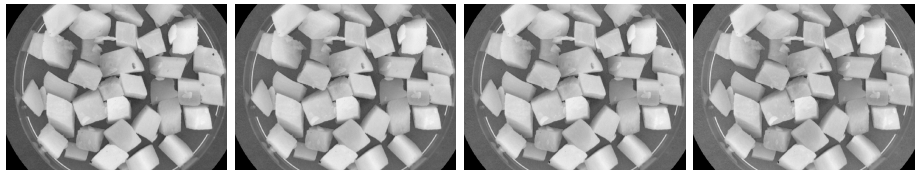
(a) Ultra blue (395nm) (b) Ultra blue/Blue (435nm) (c) Blue (450nm) (d) Blue/Cyan (470nm)



(e) Cyan (505nm) (f) Cyan/Green (525nm) (g) Amber (570nm) (h) Orange (590nm)



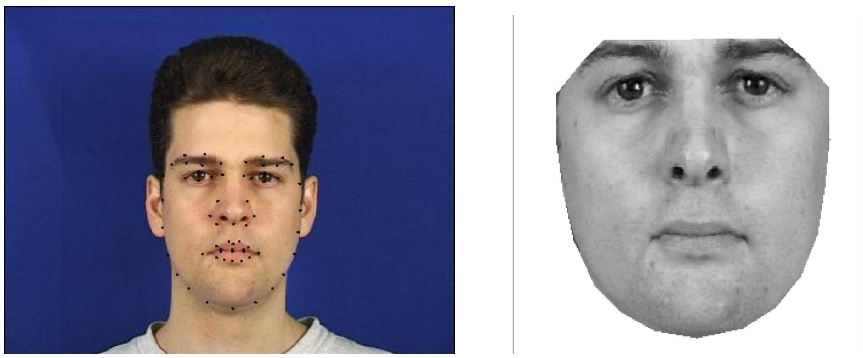
(i) Red (630nm) (j) Red (645nm) (k) Red (660nm) (l) Red (700nm)



(m) NIR (850nm) (n) NIR (870nm) (o) NIR (890nm) (p) NIR (910nm)

(q) NIR (940nm) (r) NIR (950nm) (s) NIR (970nm)

Figure 1.3: Illustration of a multi-spectral image of carrot samples. Each subfigure shows one of the spectral bands. As expected, the carrots clearly give a larger response for wavelengths in the yellow and red area than for wavelengths in the blue and green areas.



(a) RGB image and landmarks (b) Gray scale texture region

Figure 1.4: Illustration of one of the face color images in the XM2VTS database with superimposed landmarks which represent features in the face. The second image illustrates the region of interest with the color texture mapped onto a mean face region.

for the specialists. The final goal is to perform an objective classification of the three fish species using an RGB camera in the construction line; as illustrated in figure 1.6.

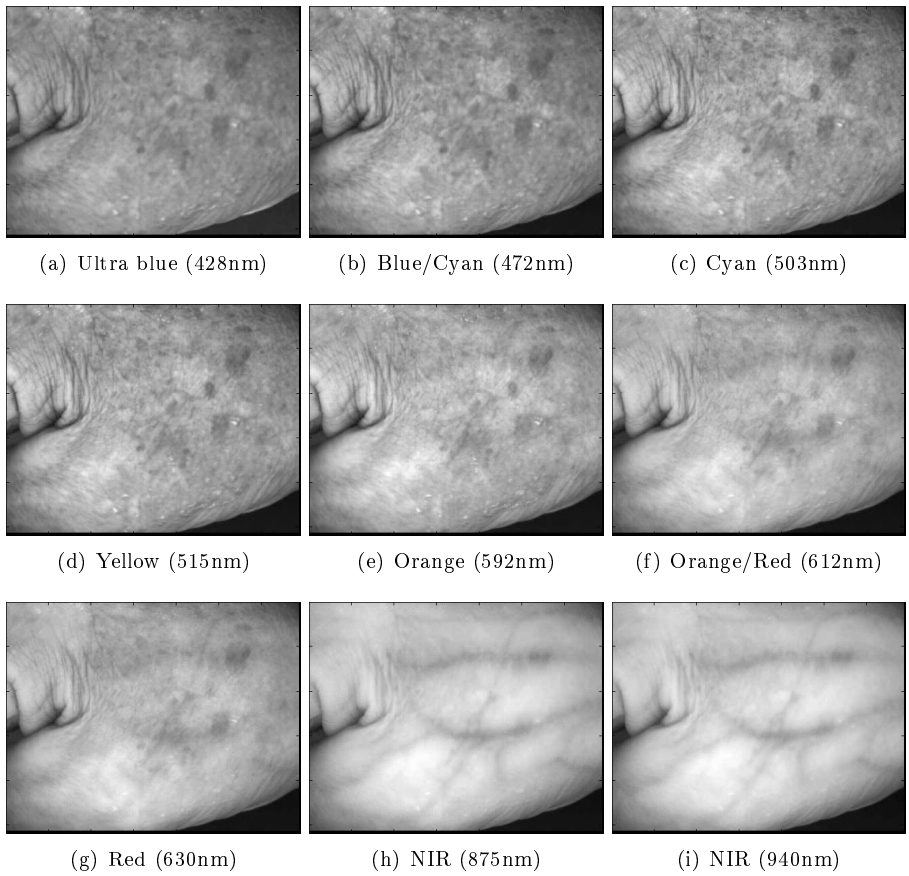
The data example here consists of standard RGB (Red, Green, Blue) color images of 108 fish from these three species. Each image contains $512 \times 768 \times 3$ color intensities; see figure 1.7. The task is to classify the species into a categorical class variable with three levels: Cod, Haddock, and Whiting.

1.1.6 Example 6 - Fungi

This example includes three different data sets. They consider classification of microbiological fungi to the species or genera level.

The first data set consists of 36 multi-spectral images of three species of *Penicillium* fungi. The *Penicillium* genus is well known as some of its isolates are used to produce foods such as fermented cheeses and salamis, and others can produce the drug *penicillin* which is effective in fighting bacteria. The multi-spectral images are of the size $960 \times 1280 \times 18$. Some examples of *Penicillium* fungi in pseudo RGB colors are illustrated in figure 1.8.

The second data set consists of 1D nuclear magnetic resonance (NMR) spectra of fungi for classification at the genus level of three fungal genera: *Aspergillus,*



(a) Ultra blue (428nm) (b) Blue/Cyan (472nm) (c) Cyan (503nm)

(d) Yellow (515nm) (e) Orange (592nm) (f) Orange/Red (612nm)

(g) Red (630nm) (h) NIR (875nm) (i) NIR (940nm)

Figure 1.5: Illustration of a multi-spectral image of a psoriasis lesion of one of the patients. Each subfigure shows one of the spectral bands.

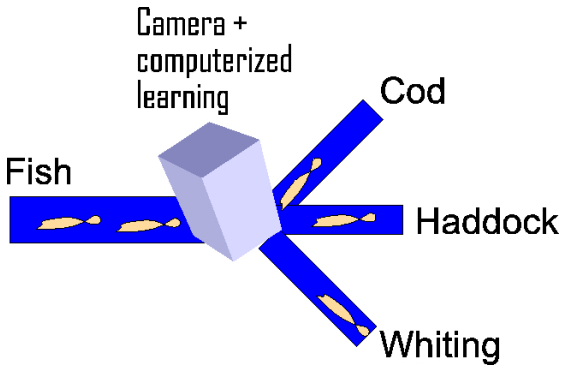
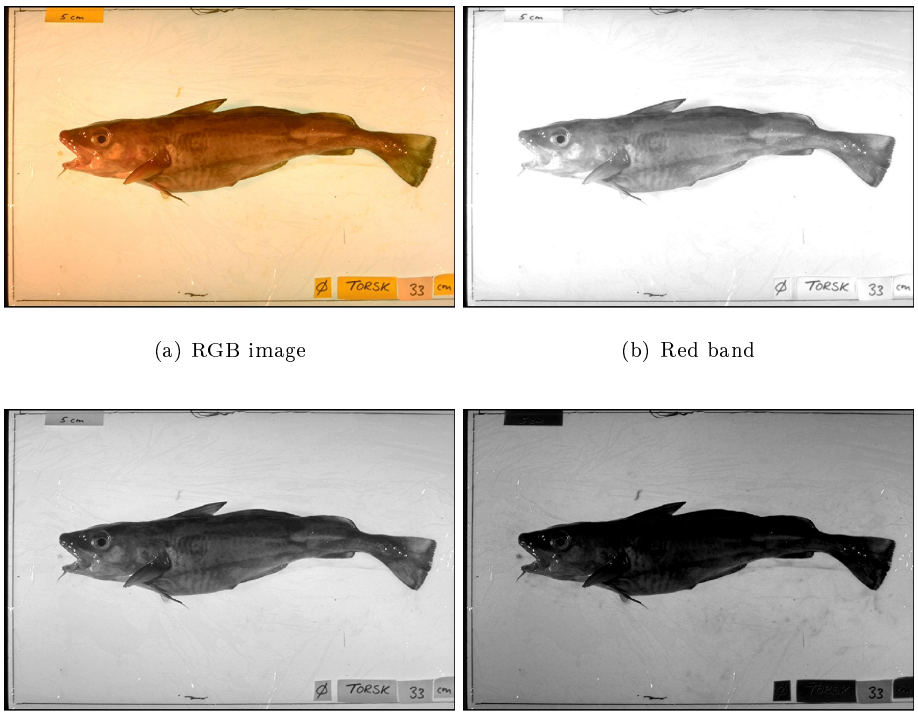


Figure 1.6: Diagram of the sorting of fish in a construction line. The grey box illustrates the camera as well as the computerized classification of the fish species which is the task considered here.



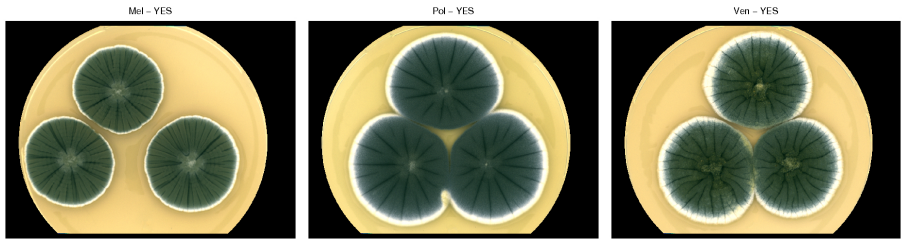
(a) RGB image

(b) Red band

(c) Green band

(d) Blue band

Figure 1.7: Illustration of one of the RGB images of a Cod fish.



(a) *P. Melanoconidium* (b) *P. Polonicum* (c) *P. Venetum*

Figure 1.8: Illustration of pseudo RGB images of three species in the *Penicillium* genus.

Neosartorya, and *Penicillium*. There are samples of various species included for each genus. For each genus there are 5, 2, and 5 species, respectively. There were 71 observations with 4-8 samples of each species. The data include information from the 950 highest peaks in the NMR spectra as features. An example of a 1D NMR spectrum is shown in figure 1.9.

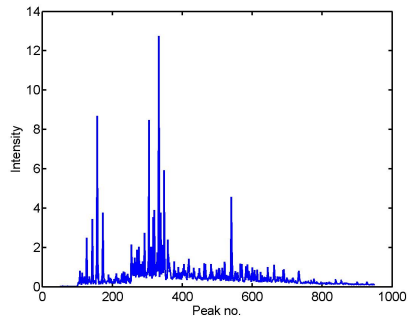
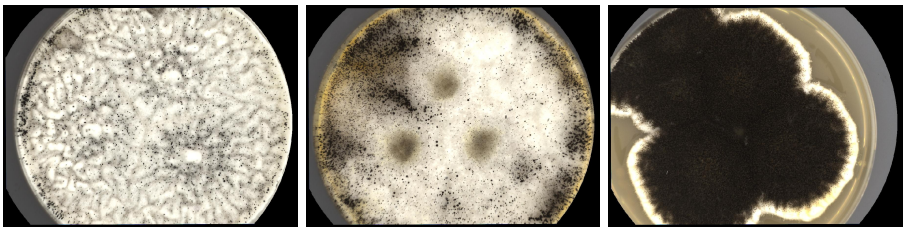
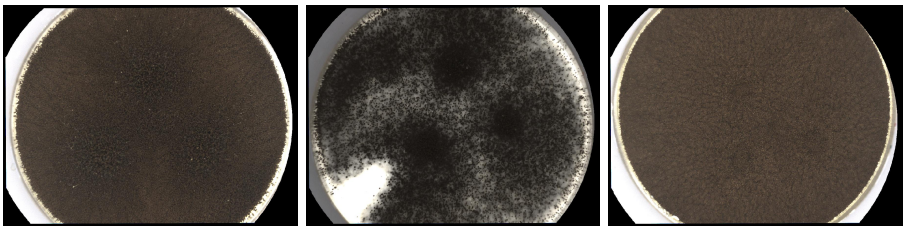


Figure 1.9: Illustration of a 1D NMR spectrum of one of the fungal samples. The spectrum includes the 950 highest peaks (frequencies).

The third data set considers classification of the two species *Niger* and *Tubingensis* of the *Aspergillus* genus. The data set consists of multi-spectral images of 33 duplicates of fungi strains grown on two different growth substrates; a total of 132 images. The images consist of $960 \times 1280 \times 18$ spectral reflectance values. Examples of pseudo RGB images of the species are illustrated in figure 1.10.



(a) *A. Niger* (b) *A. Niger* (c) *A. Niger*



(d) *A. Tubingensis* (e) *A. Tubingensis* (f) *A. Tubingensis*

Figure 1.10: Illustration of pseudo RGB images of two species in the *Aspergillus* genus.

1.1.7 Example 7 - Ear canal impressions

This example is from the hearing aid industry which before making a hearing aid for a patient models the ear canal such that they can make a hearing aid that fits exactly to that individual. One of the problems for the hearing aid manufacturer is that there is no standardized way of taking ear canal impressions. Some doctors use an open mouth setting others a closed one. Furthermore, the change of the shape of the ear canal when opening and closing the mouth (for example when chewing or talking) can cause discomforts to some patients with hearing aids. Therefore, it is of importance to be able to classify an ear canal impression to an open or closed mouth setting before making the hearing aid. The data set consists of 134 ear canal impressions scanned with a laser-scanner (taken from 67 individuals with one from each setting). Each scan of an impression results in 4356 points in 3D; resulting in a 13068 dimensional vector to represent each sample. An example of such an ear impression can be seen in figure 1.11. In the same study ear canal impressions were also taken when the individual turned his or her head to the left, but only for 42 of the individuals.

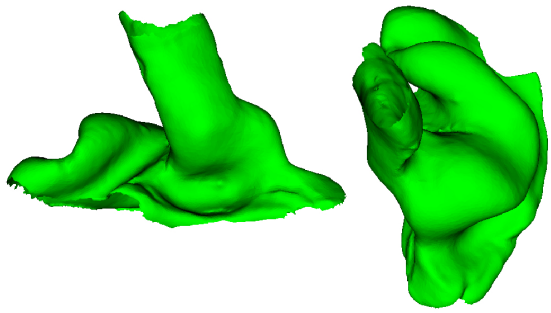


Figure 1.11: Illustration of the scanning of an ear impression shown from two angles.

1.1.8 Example 8 - Microarrays

Another popular example of large p, small n problems is gene expressions from various types of biological material. Here we consider gene expressions of several cancerous and healthy tissues measured using DNA microarrays. The idea is that an array is constructed which for a number of specified genes measures the expression level of each gene for the given tissue samples. This thesis includes three microarray data sets concerning classification between different cancer types or between healthy and cancerous tissues, respectively. The three data sets contain 12600, 3226 and 2000 probe sets (gene expressions), respectively, in the microarrays. And the respective numbers of samples are 248, 22 and 62. The three data sets are labelled into 6 leukemia subtypes, healthy and cancerous tissues, and again healthy and cancerous tissues, respectively. For the three examples it is of interest to be able to classify between the different tissue types. An example of one of the microarray data sets is illustrated in figure 1.12 where the color represents the expression level, the rows are the genes, and the columns are the samples.

These examples are used in the chapters 4 (paper A) and 6 (paper C).

1.2 What this thesis does and does not include

The present thesis considers analyses of data which have relatively more variables than observations ($p \gg n$), also called large p, small n problems. The data examples are described in the previous section and are restricted to continuous outputs (regression problems) or categorical outputs (classification problems).

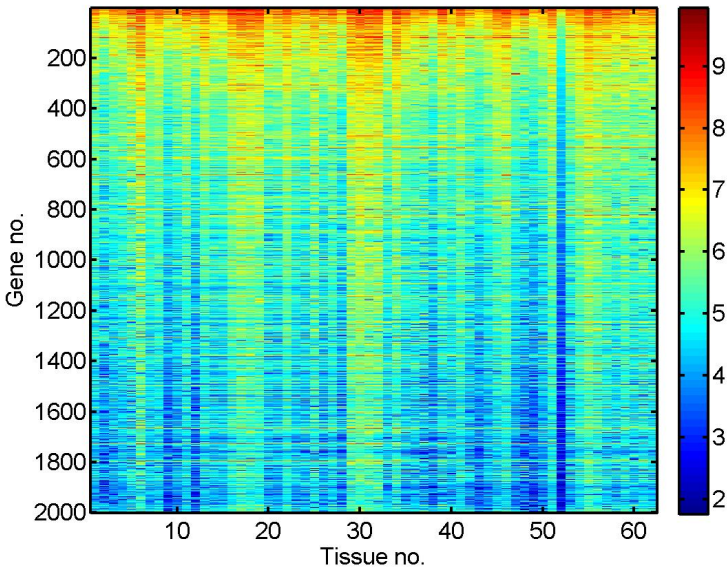


Figure 1.12: Expression profile of one of the microarrays. Red corresponds to high gene expressions and blue to low gene expressions. The colormap is made of the logarithm of the expression values.

The research in this thesis has been limited to supervised analysis where *a priori* knowledge, knowledge of class labelling or measurements of a continuous output variable, is used to train the models. This is in contrast to unsupervised analysis where *a priori* knowledge is unknown or disregarded and patterns in data are modelled without knowing *the ground truth*.

In the next chapters we will see why dimension reduction is of great importance in high dimensional problems with a scarce sampling. Here, we will note that it is necessary and that there are various ways to perform dimension reductions such as:

Feature selection or extraction For these methods features are either selected based on maximizing or minimizing an information criterion or they are extracted which means that we exclude the features which are redundant, irrelevant, or do not contain sufficient information. Examples of such methods are *forward selection* and *backward elimination*, or the *forward stagewise* technique which is a combination of the just named selection and extraction techniques.

Regularization of parameters In this approach, a regularizing term on the parameter estimates which penalizes for the complexity of the model is added to the loss function of the model. Well known regularization methods are *ridge regression* and the *lasso* (least absolute selection and shrinkage operator). The regularizing terms can also be seen as priors on the distribution of the parameter estimates.

Projections to lower dimensions Some of these methods are also known as latent variables methods or decompositions of the feature space. Most of the other dimension reduction methods can be seen as special cases of such low-dimensional projections under the name of *manifolds*. Two well known latent variable methods are *principal component analysis* and *partial least squares*.

Clustering of features Here features are clustered together and only the modes of each cluster are used in the modelling procedure. Such an approach can in particular be strong when there are strong correlations between features and a natural clustering exists. In this way more robust estimates can be obtained using the modes of each cluster as features, i.e. averaging over features.

Structuring parameter estimates Here the main idea is to structure parameter estimates according to the underlying structure of data such as spectral data which possess correlations between spectral bands. In line with this is also approximations of parameters by simpler models. Often when we are in high dimensions, simple linear models will suffice to

model data even though the sampled data has a different structure than the assumed Gaussian distributions. Such approaches are e.g. discussed in Berge (2007).

In this thesis the main focus is on dimension reduction using regularizations. Methods using other dimension reduction techniques are mainly included for comparisons. However, the point is not to choose one approach over another. The choice of dimension reduction should always be seen in the light of the problem at hand. If we look for the maximal variance in data such as if we want to describe the natural variations of human faces, then a principal component analysis (PCA) is a good choice. On the other hand if we want to classify between males and females the PCA may no longer be the best approach as the maximal variance in the faces may not be that between males and females.

One of the big advantages of using parameter regularization is that dimension reduction in this setting becomes part of the modelling technique and thus if we choose a good model for the data at hand there is no need to perform any preprocessing or find an appropriate selection technique which match the chosen modelling technique. Further advantages or disadvantages depends on which regularizations we choose to include in the model.

1.3 Reading guidelines

In this section I will give some guidelines on how to read the present thesis.

Methodology and applications parts The thesis consists of two parts: A theoretical part and a part on applications. The theoretical part consists of a chapter on theoretical considerations for the types of problems which are regarded in this thesis, a chapter on the theoretical methods which are the base of some of the further theoretical developments in the papers, and finally a chapter for each of the theoretical papers included in the present thesis. The second part of the thesis consists of a chapter for each paper which concerns applications of the methods described in the previous part and one chapter with unpublished results.

Papers An extended abstract is included as a chapter for each paper which makes part of this thesis. The abstracts summarize the papers but also put them into the context of the present thesis. The papers themselves have been placed in the appendices. This was done such that it is easier to get an overview of the work this thesis includes and decide on which papers may be of interest to the individual reader for further studies.

Reading flow The intension when writing the thesis was that it should be read from the beginning to the end and that the papers included in the appendices then could be read separately according to the reader's interests. However, I have included references from the paper chapters to the methodology so that if a paper is read without reading the thesis first it becomes easier to find relevant information if necessary. The index may also be used in this context.

Notation The mathematical notation used throughout the thesis have been adapted form the various papers such that the basic theories matches the amendments. As the notation style is different from journal to journal there may be some inconsistencies. However, it should always be clear from the context of either the theoretical section or the paper how the notation is used.

Index The index which is placed in the very end of the thesis was made such that main concepts and definitions easily can be found in the thesis. Abbreviations are also included in the index.

Part I

Methodology

This part includes an introduction to the methodology and a theoretical chapter with an introduction to the basic methods which were utilized or further developed in the included papers. Sequently, this part includes scientific papers which include methodical or algorithmic novelties or insights.

Chapter 2

Introduction to methodology

This chapter gives an introduction to the research area of this thesis: *large p, small n problems* (many features p, few observations n). It describes some of the advantages (*blessings*) and disadvantages (*curses*) for such problems.

Finally this section gives a discussion of supervised vs. unsupervised analysis and the approach in the present thesis.

2.1 The curse of dimensionality

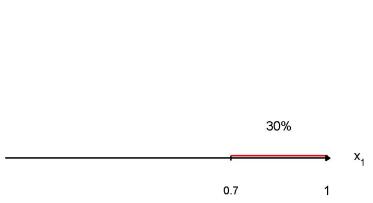
Richard Bellman introduced the term *curse of dimensionality* in 1961 to describe the problem of an exponential growth in volume when the dimensions of the input space increases (Bellman, 1961). For example, the volume of a hypercube with 100 units (sampled intervals) in each of its p dimensions has a volume of $V = 100^p$. Regarding the problems, Bellman said: "This does not mean that we cannot attack them. It merely means that we must employ more sophisticated methods."

In order to illustrate the curse of dimensionality, it is common to consider data which is uniformly distributed inside a p-dimensional unit hypercube, or a p-dimensional unit ball; see e.g. Hastie et al. (2009). In the following, two formulae

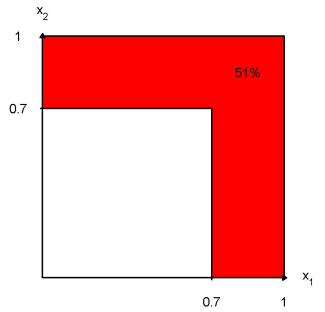
which describe the relationships of such curved or sphered data will be revised. The first one gives the expected edge length needed to capture a fraction r of the data in the unit hypercube:

$$e_p(r) = r^{\frac{1}{p}} \quad . \tag{2.1}$$

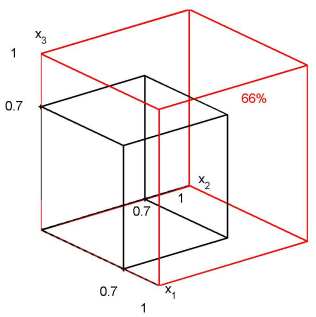
Figure 2.1 and 2.2 (a) illustrate that when p increases, it is necessary to cover a larger and larger range of each variable in order to capture the same fraction of data. This makes it infeasible to perform local estimations in high dimensions, as local estimations either become global, or if we use fewer observations in the local neighborhoods, the variance of our estimates increases.



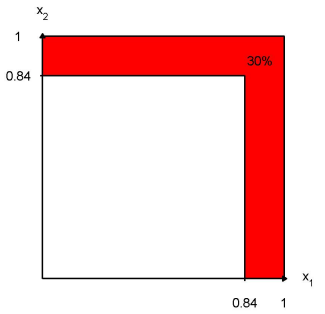
(a) 1D; covering 70% of x_1, and 70% of data.



(b) 2D; covering 70% of x_1 and x_2, and 49% of data.



(c) 3D; covering 70% of x_1, x_2 and x_3, and 34% of data.



(d) 2D; explaining 70% of data, and covering 84% of x_1 and x_2.

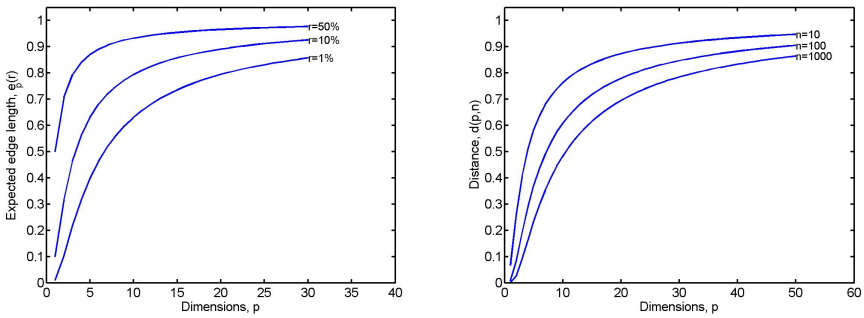
Figure 2.1: Illustration of curse of dimensionality through (2.1) with cubed data.

Related to this issue is that the distance between neighboring points increases. The second formula we revise gives the median distance from the center of the

unit hyperball to the closest data point:

$$d(p,n) = \left(1 - \frac{1}{2}^{\frac{1}{n}}\right)^{\frac{1}{p}} \quad . \tag{2.2}$$

Figure 2.2 (b) illustrates that the median distance from the center to the nearest point quickly increases for increasing p to more than halfway to the boundary ($d > 0.5$), even for large sample sizes. This means that more than half of the data points are closer to the boundaries than to any other data point. Therefore, interpolations are most often infeasible in high-dimensional spaces and rather become extrapolations which often have less predictability.



(a) Range of variables

(b) Distance to closest point

Figure 2.2: Illustration of the curse of dimensionality through (a) the range of each variable we need to cover as a function of the dimensions when we want to describe 1%, 10%, and 50% of data, respectively; (2.1), and (b) the median distance to the closest point as a function of the dimensions with sample sizes $n = 10, 100, 1000$, respectively; cf. (2.2).

Furthermore, the sampling size required to obtain a dense sampling in the p-dimensional space grows with the power of the dimensions; $N_p = N_1^p$ (where N_1 is the sample size required to obtain a dense sampling in the 1-dimensional space). In many cases it is not possible to collect such massive amounts of observations. Either because it is too expensive, too time consuming, or because the samples of interest are rare. Finally, it can be conceptually difficult to interpret high-dimensional spaces. All of this makes it attractive to reduce the dimensions, either directly by some sort of feature selection or at least the effective dimensions by some sort of parameter regularization or latent variables.

2.2 Blessings of dimensionality

In 2000 David Donoho gave a presentation on the 21st century blessings and curses of dimensionality at the conference of *Math Challenges of the 21st Century*; Donoho (2000). He described how several blessings follow the curse of dimensionality. The first blessing comes from probability theory, and assumes that there are many similar (highly correlated) features in the high-dimensional data which one can average over to obtain better estimates. This is in general the case for the examples presented in this thesis. For example, the spatial coherence in the images result in spatial correlations between pixels both within a spectral band, but also across spectral bands. The second blessing comes from the central limit theorem which says that there is an underlying limit distribution which is approached as the number of dimensions go to infinity. This means that data lie on a low-dimensional manifold. The third of the blessings arises when measurements are taken from an underlying continuous process such as spectra or images. For such data the underlying structure often gives an approximate finite dimensionality to the sampled data despite the high dimensionality, i.e. the underlying structure is often of a simple structure which can be exploited. This again means that data lie on a low-dimensional manifold.

In particular for the examples which use multi-spectral images these blessings should be exploited. In this thesis we have used summary statistics to obtain fewer and more robust features which describe the objects of interest in such multi-spectral images. More details on this matter can be found in chapter 8 and paper E. Furthermore, for all data examples we assume that the data lie on low-dimensional manifolds, and thus we seek low-dimensional, simple structures in data.

2.3 Supervised vs unsupervised analysis

Unsupervised analysis is a term used for methods which do not use *a priori* information about an output label of data. These methods look for patterns in data described by some criterion such as maximum variance, maximum correlation or minimum distance between input variables. The unsupervised methods are powerful as they can reveal patterns of information in data which we had no prior knowledge of.

Supervised analysis is the term used for methods which use *a priori* information via an output label for training data. This means that we have given continuous measurements or class labels for a training set of data, and we build the model

based on this information. The supervised analysis is powerful when good a priori information exists. However, one should note that it is important that the model generalizes to new data instances and not only describes the training data (see more in the section on overfitting; section 3.2.1).

In the last decade there has been emphasis on *semi-supervised learning* where unlabelled data is used to improve predictions of a model trained with labelled data. A semi-supervised setting can for example be performed in a deterministic annealing setting as in e.g. Leistner et al. (2009) or with a penalization term on varying class labels in high density areas as in Fergus et al. (2009). In general, an updating of a model with unlabelled data improves performance and it is in particular useful when labelled samples are expensive (e.g. performed manually), and in online settings where unseen/unpredicted data can be used to adapt a given model. However, this thesis will only contain the supervised learning setting. On the other hand, it should be noted that these methods could be extended to use in semi-supervised settings as well. Actually, for some of the methods, e.g. support vector machines, such learning structures have already been developed, for more information see e.g. Belkin et al. (2006).

Chapter 3

Basic methodology

This chapter describes the basic methodology which the included papers build on or in other ways make use of for statistical analyses. The idea is to give a brush up on the methods and a good intuition about their usage and a few tips on calculation issues. It is not intended to be a thorough statistical text book and rather provides useful citations for such where further details may be studied. There are several related methods and also various traditional methods which have been used for comparisons in the papers included in the appendices, but are not included in this section. To get a full overview of related methods I refer to the citations used in the papers and to for example the following text books: Rencher (2002); Hastie et al. (2009); Duda et al. (2001); Bishop (2006), or other overview text books in the fields of statistics, statistical learning, pattern recognition and machine learning.

The papers in this part of the thesis build on the methods listed in this introductory chapter. The methods described here are classical statistical methods which work well when the number of observations is larger than the number of features ($n > p$). The first section is on ordinary least squares regression which in general is used when the dependent predictor variable is continuous. Section two describes why dimension reduction is important for large p, small n problems and introduces the terms of overfitting and bias-variance trade off. Section three, four and five describe the three classification methods: Linear discriminant analysis, mixture discriminant analysis, and support vector ma-

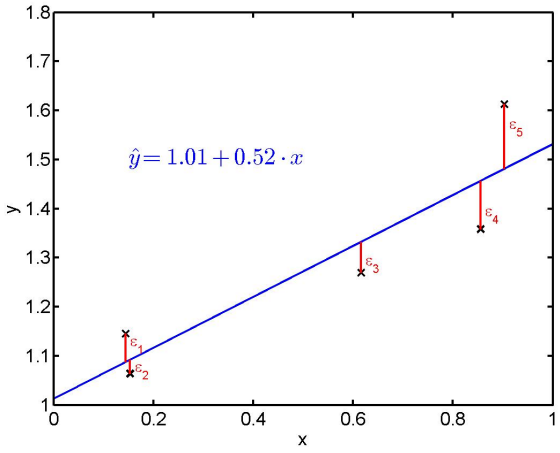
chines. Section six describes two well known regularizations of the parameter estimates (the ℓ_1- and ℓ_2-norm) and their differences, and the advantages of using them jointly. Section seven discusses cross-validation and its importance in supervised analysis.

3.1 Ordinary least squares regression

Ordinary least squares (OLS) regression is a widely used statistical tool which provides linear models that link p input variables x to an output variable y using a parameter vector β of size $p \times 1$. Assuming that y is centered ($\sum_{i=1}^{n} y_i = 0$) or the intercept is included in X such that $X = [1, x_1, x_2, ..., x_p]$, the linear regression model can be written as

$$y = X\beta + \epsilon \quad , \tag{3.1}$$

where y is an $n \times 1$ vector of the outputs, X is an $n \times p$ matrix of the inputs (or $n \times (p+1)$ if the intercept is included), and ϵ is an $n \times 1$ vector of the model errors, also called residuals or noise. It is seen that the residuals are given by $\epsilon = y - X\beta$. Geometrically the residuals are the distances from the observed outputs y to the estimated outputs $\hat{y} = X\hat{\beta}$; see figure 3.1. Additional



(a) Range of variables

Figure 3.1: Illustration of a linear regression model in one dimension.

common assumptions for the linear regression model are that the residuals are independent and normally distributed, i.e. $\epsilon \in N(0, \sigma^2 I)$.

For a statistical description of linear regression models the text book Rencher (2002) is useful. For a description which is less theoretical, and more in the alley of data mining and machine learning, see Hastie et al. (2009).

In general the total error of the model is defined using the residual sum of squares (RSS), which is given by

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 = (y - X\beta)^T(y - X\beta) \quad . \tag{3.2}$$

The residual sum of squares is a quadratic convex function and therefore the minimum is obtained where the differential is zero:

$$\frac{\delta RSS}{\delta \beta} = X^T(y - X\beta) = 0 \quad , \tag{3.3}$$

which are also called the *normal equations* (recall that this is a system of equations as y and β are vectors and X is a matrix). If X^TX has full rank (is nonsingular) then the normal equations have a unique solution

$$\hat{y} = (X^TX)^{-1}X^Ty \quad . \tag{3.4}$$

If X^TX does not have full rank it is common to use a pseudo inverse to replace $(X^TX)^{-1}$. For more details on pseudo inverses, also called generalized inverses, see e.g. Ben-Israel and Greville (2003).

For linear regression methods it is common to use the mean squared error (MSE) as a measure of how good the predictions are. The MSE is defined as

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 = \frac{1}{n}(y - X\beta)^T(y - X\beta) \tag{3.5}$$

Note that this is just a scaled version of the RSS. It is also useful to plot the predicted values against the true output $(\hat{y}, y)$ which also gives the ability of looking for trends and possible nonlinearities. For a detailed list on how to check linear regression fits one may address Rawlings (1998); Ersbøll and Conradsen (2003). Here, we will just note that the basic idea is to test if the assumptions of the linear regression model hold.

3.1.1 Calculation issues

If X^TX has full rank but is badly conditioned, i.e. the columns are close to linear dependent then X^TX should not be inverted directly as this will result

in instable solutions. Instead $\hat{\beta}$ may be calculated using a QR matrix decomposition of X. This is common in numerics and can be done with the following Matlab command $\hat{\beta} = X\backslash y$. In fact, this Matlab command can handle both over- or underdetermined systems of equations and always gives more accurate solutions than explicit inversions of X^TX. That is because X^TX has a condition number that is twice as large as that of X, and consequently the precision of the solution using $(X^TX)^{-1}$ will be only half the precision of the solution using X^{-1} . If the system is underdetermined, where X^TX does not have full rank, then the QR decomposition is used to calculate the effective rank and truncate the solution accordingly. Note, that this in general is not the same as using a pseudo inverse to replace $(X^TX)^{-1}$. For more information on QR decompositions and numerical methods for solving linear systems of equations see Elden (2007).

3.1.2 Non-linearities and random forests

When using a linear regression model, we should keep in mind that there may be non-linear effects which we will not encounter. To check if there are non-linearities we may run for example a random forest analysis and check if prediction drastically improves (Breiman, 2001). In brief, random forests is a collection of independent predictor trees which are based on a random selection of features to split each node. Some of the advantages of random forests are that they are fast, they are robust to noise, and they are non-linear. But also residual vs. independent predictor plots can be useful, although with a vast number of independent predictors this can become cumbersome.

3.2 Dimension reduction

This section illustrates why dimension reduction is necessary in large p, small n problems and in the process introduces the terms of *overfitting* and *bias, variance trade off*.

3.2.1 Overfitting

We can never do worse with regards to classification rates or error measures on the training data by adding a new dimension to our model (as we can always set the parameter for the new variable or component to zero), but we do risk

to *overfit*. Overfitting means that we fit our model well to the training data (the data that is used to train or fit the model), but that the model does not generalize, i.e. it does not predict new data well (often called validation or test data). We can choose to overcome or address this problem by e.g. cross-validating[1] and testing to make sure our model generalizes to new inputs (fits training and testing data equally well). On the other hand, we also need to make sure we don't underfit, i.e. we still want a solution which is sufficiently rich to answer the questions at hand. The next section illustrates, by use of an example, what happens when we overfit or underfit our model in terms of predictability.

3.2.2 Bias, variance trade off

The MSE, introduced in the section on linear regression, can be written as:

$$\begin{aligned} MSE(x) &= E_{X_{train}}[f(x) - \hat{y}]^2 && (3.6)\\ &= E_{X_{train}}[\hat{y} - E[\hat{y}]]^2 + \left(E_{X_{train}}[\hat{(y)}] - f(x)\right)^2 && (3.7)\\ &= \mathrm{Var}_{X_{train}}(\hat{y}) + \mathrm{Bias}^2(\hat{y}) \quad , && (3.8) \end{aligned}$$

where $f(x)$ is the true underlying function, $\hat{y}$ the prediction of y for input x, and $E_{X_{train}}[\cdot]$ denotes the expectance given the training data. This is called the *bias-variance decomposition* of the MSE. We see that the bias is the difference in the expected value of the prediction of y and the true value of y; this can also be interpreted as an offset in the prediction model in comparison with the true model. The variance of the prediction tells us about the expected difference between the prediction and the expected prediction. Note, that the bias-variance decomposition can be made for any loss function that we choose. In the following we look at two examples of linear regression models and the trade off we make between bias and variance when we determine the dimensionality (or complexity) of the model.

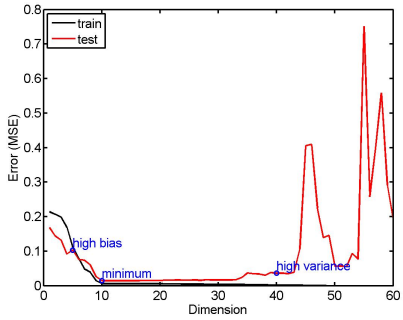
We consider an example with 60 uniformly distributed random variables $x_1, ..., x_{60}$ ($p = 60$) for the linear model

$$f(x) = 1 + 0.5 * (x_1 + \cdots + x_{10}) + \epsilon \tag{3.9}$$

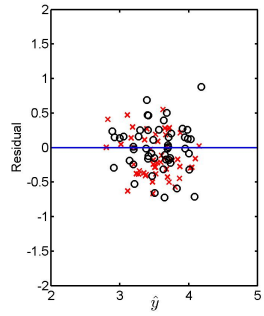
with noise $\epsilon \in N(0, 0.5^2)$. Note that only 10 of the variables actually contribute to the model. We generate an outcome of 50 observations ($n = 50$) from the model. Figure 3.2 and 3.3 illustrate the MSE as a function of the dimensions for two feature selection algorithms utilized on an outcome of (3.9). Note, that

[1] Cross-validation will be explained in 3.7

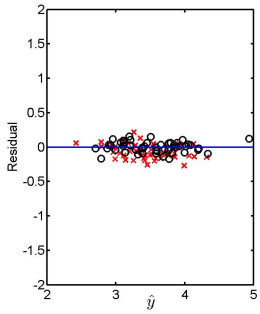
in many textbooks such $(dimensions, MSE)$-plots are illustrated using smooth curves; e.g. (Hastie et al., 2009, Figure 2.11). These are often averages over several outcomes of a model, or made for a simple and more controlled setting without noise. Here, we illustrate how such figures may look for a single outcome of a model with noise. This results in more wiggly curves. At each step i on the dimension axis in figure 3.2 we make an OLS fit including the first i variables; $x_1, ..., x_i$. This means that the first 10 variables which are truly in the model get selected first, and subsequently the variables have a random correlation with the output y of the model. The minimum of the test error MSE_{test} is when



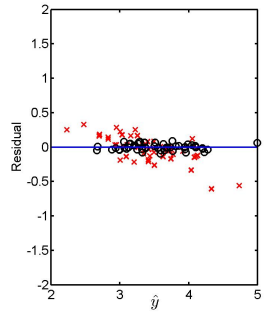
(a) Error vs. Dimension



(b) Ex: High bias



(c) Ex: Minimum



(d) Ex: High variance

Figure 3.2: Illustration of a linear regression model with varying bias and variance trade off. First, the MSE vs. the number of variables (dimensions) included in the model. Here, the variables are selected one by one according to their index number. Furthermore, three examples are depicted of the residuals vs. the predicted output.

10 dimensions are included. Often the intersection between the MSE and two times the standard error of the cross-validation is used to choose the dimensions

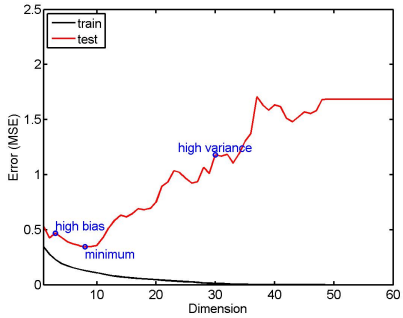
of the model. As a rule of thumb it is a good idea to keep the dimensions as low as possible when modelling $p \gg n$ problems, look for the first *kink* where the training and test errors flatten out.

In figure 3.3 the same example is illustrated but where the variables are included according to the highest partial correlation with the dependent variable y and conditioned on the input variables already included in the model. The first 10 selected variables were: $x_5, x_4, x_7, x_6, x_9, x_{30}, x_{49}, x_2, x_3, x_{38}$. Note, that only 7 of these are in the true model. The minimum of the MSE_{test} is now when 8 variables (dimensions) are included. If forward selection with a significance level of 5% was used only 6 variables were included (5 of these are in the true model). This illustrates well that even with various variable selection criteria we may include features which are actually irrelevant for the true underlying model. This is in particular an issue when n is small. Consider the example of generating thousands and millions of random features for a limited number of observations, then eventually one of them will give a high correlation with the observed output. This is an issue in $p \gg n$ problems we must be aware of as we can never be sure to entirely eliminate this fact unless we obtain more samples.

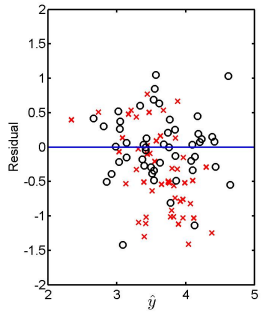
In general the MSE is dominated either by the bias or by the variance when the dimensions are high, this is an effect of the curse of dimensionality; see (Hastie et al., 2009, section 2.5, figure 2.7 and figure 2.8). This is why one of the approaches to solve the curse of dimensionality is to trade off some variance for a reduction in the bias or vice versa trade off some bias for a reduction in the variance, and thereby obtain a lower MSE. Whether it is the bias or the variance that dominates the MSE depends on the complexity of the model, as it has been illustrated in this section. The complexity of the model can e.g. be penalized with a regularization term; see section 3.6.

3.3 Linear discriminant analysis

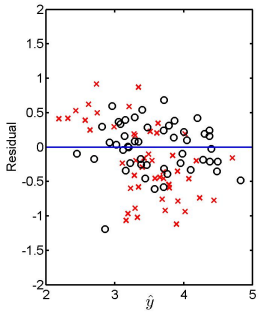
The idea behind discriminant analysis is to minimize the variance within each class and maximize the variance between classes. The method was first proposed by Ronald Fisher in 1932 to classify three species of the iris flower; Fisher (1936). In linear discriminant analysis the functions which separate the classes are linear. The method can also be extended to quadratic separating functions. For k classes, we maximize the between-groups sums of squares, $\Sigma_B = \sum_{j=1}^{k} (\mu_j - \mu)(\mu_j - \mu)^T$ (where μ is the mean of all groups) relative to the within-groups sums of squares, $\Sigma_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \mu_j)(X_{ij} - \mu_j)^T$



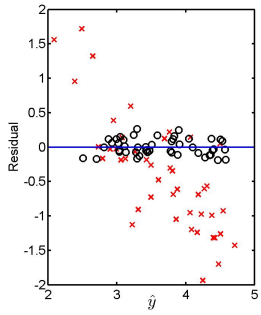
(a) Error vs. Dimension



(b) Ex: High bias



(c) Ex: Minimum



(d) Ex: High variance

Figure 3.3: Illustration of a linear regression model with varying bias and variance trade off. First, the MSE vs. the number of variables (dimensions) included in the model. Here the variable selection is made according to the partial correlation between y and x. Furthermore, three examples are depicted of the residuals vs. the predicted output.

(where μ_j is the mean of the j^{th} class)

$$\arg\max_{\beta_j} \beta_j^T \Sigma_B \beta_j \tag{3.10}$$

under the orthogonality constraint

$$\beta_j^T \Sigma_W \beta_l = \begin{cases} 0 & l = 1, ..., j-1 \\ 1 & l = j \end{cases} \quad , \tag{3.11}$$

to find the discriminating directions β_j, $j = 1, ..., k-1$. This is also known as Fisher's criterion.

The same classifiers can also be expressed using a Bayesian framework[2]. The Bayes classifier simply states that we classify to the most probable class based on the conditional distribution $P(G|X)$ (the probability of class G, given X). Applying Bayes theorem[3] gives us

$$\begin{aligned} P(G = j|X = x) &= \frac{P(X = x|G = j)P(X = x)}{P(G = j)} \\ &= \frac{f_j(x)\pi_j}{\sum_{i=1}^{k} f_i(x)\pi_i} \quad , \end{aligned} \tag{3.12}$$

where π_j is the prior probability of class j, and $\sum_{j=1}^{k} \pi_j = 1$.

Assuming normally distributed populations $\pi_j \simeq N(\mu_j, \Sigma_j)$, $j = 1, ..., k$, we have the class-conditional density

$$f_j(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1}(x - \mu_j)\right) \quad . \tag{3.13}$$

With a common covariance matrix Σ (the classes are assumed to have the same dispersion) we get linear discriminant functions, and with a separate covariance structure for each class, the discriminant functions become quadratic. The linear discriminant functions are given as

$$\delta_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j + \log \pi_j \quad . \tag{3.14}$$

Also assuming equal class priors the *linear discriminant functions* are

$$\delta_j(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2}\mu_j^T \Sigma^{-1} \mu_j \tag{3.15}$$

[2]Bayesian statistics give analyses where probabilities are seen as a quantification of uncertainty. In contrast, classical or frequentist statistics interpret probabilities as frequencies of random, repeatable events. For more information on these differences, see e.g. Bishop (2006).

[3]$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

$j = 1, ..., K$ The decision rule is

$$G(x) = \text{argmax}_j \delta_j(x) \tag{3.16}$$

and the linear decision boundary between two classes i and j is

$$x^T \Sigma^{-1}(\mu_i - \mu_j) - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \frac{1}{2}\mu_j^T \Sigma^{-1}\mu_j \quad . \tag{3.17}$$

For LDA it is common to use the frequentist maximum likelihood (ML) estimates $\hat{\Sigma}$ and $\hat{\mu}$ based on observed data to compute classification boundaries and decisions.

Figure 3.4 illustrates the linear discriminant analysis of three classes with

$$\begin{aligned} \Sigma &= \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 2 \end{bmatrix} \\ \mu_1 &= (2, 4.5) \\ \mu_2 &= (0, 0) \\ \mu_3 &= (4, -1.5) \quad . \end{aligned}$$

Both the theoretical (based on the true Σ and μs) and the estimated (based on the ML estimates $\hat{\Sigma}$ and $\hat{\mu}$s) linear discriminant functions are shown. The ML estimates with 30 observations in each class ($N_c = 30$) are

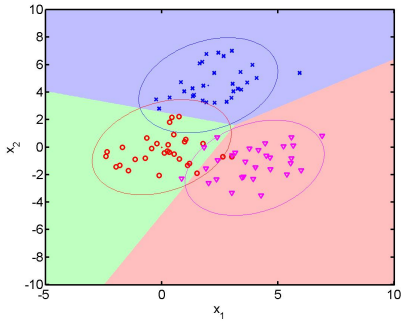
$$\begin{aligned} \hat{\Sigma} &= \begin{bmatrix} 1.9178 & 0.4614 \\ 0.4614 & 1.3793 \end{bmatrix} \\ \hat{\mu}_1 &= [2.2935, 4.6769]^T \\ \hat{\mu}_2 &= [0.0865, -0.2733]^T \\ \hat{\mu}_3 &= [3.9290, -1.1216]^T \quad . \end{aligned}$$

And with 100 observations in each class ($N_c = 100$) they are

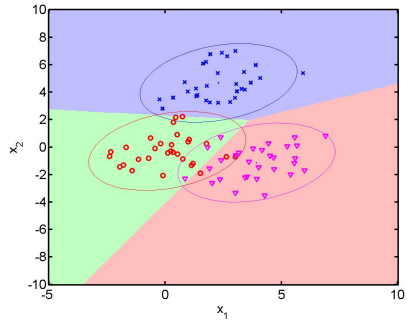
$$\begin{aligned} \hat{\Sigma} &= \begin{bmatrix} 1.6302 & 0.3768 \\ 0.3768 & 2.0293 \end{bmatrix} \\ \hat{\mu}_1 &= [2.1349, 4.5891]^T \\ \hat{\mu}_2 &= [0.24864, -0.0793]^T \\ \hat{\mu}_3 &= [3.7698, -1.5863]^T \quad . \end{aligned}$$

Notice that the estimates get closer to the true values as we have more evidence (more observations are observed). In figure 3.4 we also show the 95% confidence ellipses based on the true and estimated modes. The confidence ellipses have the equation

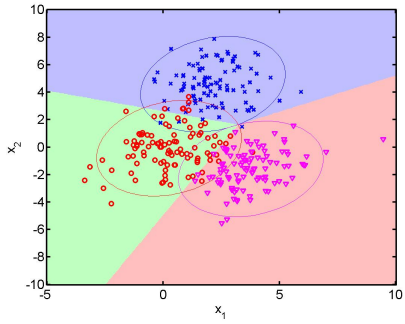
$$x^T \Sigma^{-1} x = \chi^2_{95\%}(p) \quad , \tag{3.18}$$



(a) Theoretical distributions



(b) Estimated distributions with $N_c = 30$



(c) Estimated distributions with $N_c = 100$

Figure 3.4: Theoretical and estimated distributions and decision rules for LDA. The ellipses illustrate the 95% confidence ellipses of the distributions.

where $\chi^2_{95\%}(2) \approx 6$.

In general LDA gives good classification rates (Hand, 2006). For $p \gg n$ problems there is in general no reason to add complexity by using e.g. quadratic discriminant analysis. However, we should keep in mind that the underlying distributions may be different and more complex than the assumptions allow for. One way to check for non-linearities may be to run random forests which also can be used for classifications. In section 3.4 we will get back to another way of taking into account non-linearities, namely by modelling each class as a mixture of Gaussians.

In chapter 6 (paper C) we have extended the linear discriminant analysis to a sparse discriminant analysis which works for large p, small n problems.

3.3.1 Reduced-rank LDA

Here, we will illustrate that the Bayesian approach to the LDA in (3.17) can be performed in a $k-1$-dimensional subspace as well as the Fisher approach in (C.1). This means that even when $p > k$ LDA provides low dimensional views of data. First, consider two classes in two dimensions as illustrated in figure 3.5. From (3.17) we see that only the projection of x onto $\Sigma^{-1}(\mu_1 - \mu_2)$ matters for the classification in LDA. This means that we can project data orthogonal with respect to Σ^{-1} onto the linear decision boundary; and classify a point according to the nearest centroid μ_i, $i = 1, \ldots, k$. This concept can be extended to more dimensions and so an attractive feature of LDA is the low-dimensional views that it provides. The computation of these low-dimensional projections can be performed using eigenvalue decompositions of within-group and between-group matrices as described in Hastie et al. (2009).

3.4 Mixture discriminant analysis

When a single perceptron (in this case a Gaussian distribution) is not enough to describe a class, then a mixture of Gaussians may be used to model each class. This approach is also called mixture discriminant analysis. This model extends linear and quadratic discriminant analysis to a more complex setting with non-linear boundaries. The Gaussian mixture distribution can be expressed as a

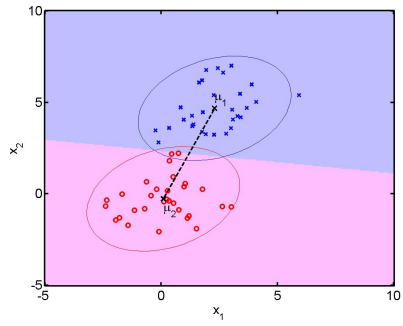


Figure 3.5: Illustration of $\mu_1 - \mu_2$ and the linear decision boundary which are orthogonal with respect to Σ^{-1}. A projection of any point parallel with the decision boundary and onto $\mu_1 - \mu_2$ gives the one-dimensional view of data ($k - 1 = 2 - 1 = 1$ dimension) which contain the classification information for the problem. The reduced rank classification is obtained as the shortest of the distances to the means (centroids) in the one-dimensional space.

linear sum of Gaussians which for the j^{th} class is given by

$$P(X|G = j) = \sum_{r=1}^{R_j} \pi_{jr} N(X|\mu_{jr}, \Sigma) \quad , \tag{3.19}$$

where π_{jr} is the mixing proportion or the subclass probability for the j^{th} class and the r^{th} subclass, with R_j subclasses for class j. The mixing proportions sum to one. For simplicity we use a common covariance matrix Σ in the notation, but this can easily be replaced with separate covariances Σ_{jr}. Finally, the mean of each subclass is denoted μ_{jr} and $N(\,)$ denotes the normal distribution.

The maximum-likelihood estimates (ML) for the parameters in the mixture of Gaussians distribution is commonly estimated using an expectation-maximization (EM) algorithm; see e.g. Hastie et al. (2009); Bishop (2006). The algorithm is as follows. *Expectation step:*

$$z_{ir} = \frac{\pi_{jr} \exp\{-\frac{(X_i - \mu_{jr})\Sigma^{-1}(X_i - \mu_{jr})}{2}\}}{\sum_{r=1}^{R_j} \pi_{jr} \exp\{-\frac{(X_i - \mu_{jr})\Sigma^{-1}(X_i - \mu_{jr})}{2}\}} \tag{3.20}$$

$$\pi_{jr} = \sum_{i \in g_r} z_{ir}, \quad \sum_{r=1}^{R_j} \pi_{jr} = 1 \tag{3.21}$$

where z_{ir} is the subclass probability of the r^{th} subclass in class i, and g_r is the

subset of observations in the r^{th} subclass. *Maximization step:*

$$\mu_{jr} = \frac{\sum_{i \in g_r} x_i z_{ir}}{\sum_{i \in g_r} z_{ir}} \tag{3.22}$$

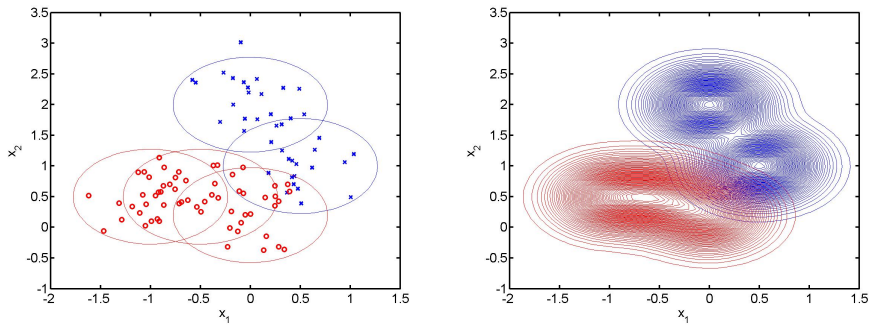
$$\Sigma = n^{-1} \sum_{j=1}^{k} \sum_{i \in g_r} \sum_{r=1}^{R_j} z_{ir}(x_i - \mu_{jr})(x_i - \mu_{jr})^T \quad . \tag{3.23}$$

The two steps are alternated until a convergence is seen in either the parameter estimates or the log likelihood.

As an example we consider mixtures of Gaussians for two classes with the following models

$$\begin{aligned} C_1 &\sim \pi_{11} N([0,2]^T, \Sigma) + \pi_{12} N([0.5,1]^T, \Sigma) \\ C_2 &\sim \pi_{21} N([-1,0.5]^T, \Sigma) + \pi_{22} N([-0.5,0.5]^T, \Sigma) + \pi_{23} N([0,0.2]^T, \Sigma) \\ &\text{where} \\ \Sigma &= \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \\ \pi_{1j} &= \frac{1}{2}, \quad j = 1,2 \\ \pi_{2j} &= \frac{1}{3}, \quad j = 1,2,3 \quad . \end{aligned}$$

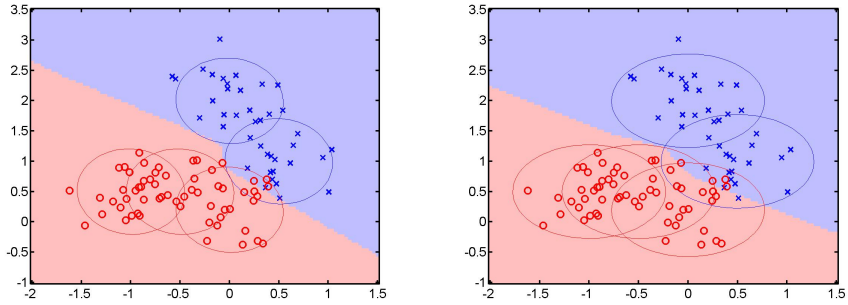
The distributions of the two classes are illustrated in figure 3.6.



(a) Confidence ellipses (b) Contours

Figure 3.6: Theoretical distributions and samples from two groups modelled by mixtures of Gaussians. The ellipses illustrate the 95% confidence ellipses of each of the normal distributions added to obtain the mixture distributions. The contour plots are of the probability density functions of the mixtures.

Using the expectation and maximization algorithm for ML estimates and with the true parameters of the two classes C_1 and C_2 the separating boundaries are illustrated in figure 3.7.



(a) ML parameter estimates

(b) True parameters

Figure 3.7: The decision boundaries between the two mixtures of Gaussians calculated using (a) the ML estimates of the distribution parameters (for 100 observations) and (b) the true parameters.

Recently a paper which integrates the reduced rank discriminant analysis to the EM framework and thereby reduces the dimensions for the mixture discriminant analysis has been proposed by Wu and Boyer (2009). In chapter 6 (paper C) we have developed a method which furthermore, and independently, introduces sparsity to the mixture discriminant analysis, i.e. an additional feature selection is performed via an ℓ_1-norm on the parameter estimates.

3.5 Support vector machines

The support vector machine (SVM) does not make any assumptions of the underlying distributions of the populations at hand. It is mostly used for classification and this is what we will revise here, but it can also be used for regression Hastie et al. (2009). SVM creates a linear hyperplane, which separates two classes, with focus on the observations close to the decision boundary (within a specified margin) and on the misclassifications. Multiclass SVMs exist, but we will not get further into that here; see e.g. Bishop (2006). In figure 3.8 an example with two classes in two dimensions is illustrated. The decision boundary $\phi(x) = \beta^T x + \beta_0 = 0$ is in this case a line, and in general it is a hyperplane separating the two classes in p dimensions (β is a p dimensional vector). The

margin of total size $\frac{2}{\|\beta\|}$ around the hyperplane decides which observations are considered when constructing the optimal hyperplane. Since the SVM in this way only uses a few of the observations to compute the decision boundary it is called a sparse method, and the computational cost is in general low.

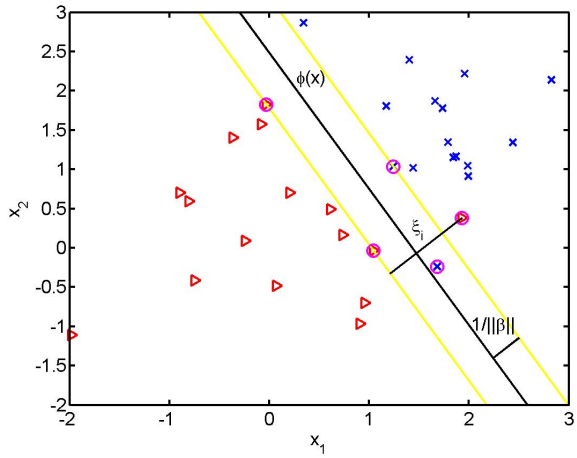


Figure 3.8: Illustration of the support vector machine. The separating hyperplane, $\phi(x)$, is the black line and the yellow lines on each side illustrate the margin of total width $\frac{2}{\|\beta\|}$. The support vectors, i.e. the points which contribute to the solution, are marked with circles, and the two classes are marked with crosses and triangles, respectively.

The support vector machine can be written as a quadratic problem with linear constraints by introducing a slack variable ξ_i for each point i. The slack variable is zero for points which are outside the margins and correctly classified, and it is the perpendicular distance from the margin to each point for all points which are inside the margins or misclassified.

$$\begin{aligned} &\arg\min_{\beta,\xi} \|\beta\|^2 + \gamma \sum_{i=1}^{n} \xi_i \\ &\text{subject} \quad \text{to}: \quad y_i(\mathbf{x}_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \quad , \end{aligned} \tag{3.24}$$

where γ is a constant determining the weight on the slack variables (positive for the misclassifications and points inside the margin) in relation to the weight on the margin size. The vector y is a class label with $y_i \in \{-1, 1\}$ for the two class case. The quadratic constraint on the parameter estimates β for the separating hyperplane has a ridge shrinkage penalization effect on β and

therefore the SVM solutions often have good generalization even for $p \gg n$ problems. However, overfitting can still be an issue, in particular when noisy, redundant, or irrelevant features are present, as all features contribute to the solutions.

A quadratic problem with linear constraints, as the SVM in (3.24), is a convex optimization problem which means that there is a global minimum and no local minima. In order to solve the quadratic programming problem we introduce Lagrange multipliers $\alpha_i \geq 0$ and μ_i for each of the constraints in (3.24). The primal Lagrangian function is given by

$$L_P(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{n} \mu_i \xi_i \quad . \tag{3.25}$$

We minimize L_P w.r.t. β, β_0 and ξ. We are maximizing the added terms $[y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)]$ and $\sum_{i=1}^{n} \mu_i \xi_i$ w.r.t. α and μ, and therefore have minus signs in front of these terms in L_P. Setting the derivatives of $L_P(\beta, \beta_0, \xi, \alpha, \mu)$ with respect to β, β_0 and ξ equal to zero, we get the following three conditions

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{3.26}$$

$$0 = \sum_{i=1}^{n} \alpha_i y_i \tag{3.27}$$

$$\alpha_i = \gamma - \mu_i \tag{3.28}$$

and we also have the positivity constraints $\alpha_i, \mu_i, \xi_i \geq 0$. We use these conditions to eliminate β, β_0 and ξ from L_P and obtain the dual representation of the SVM:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{3.29}$$

which we maximize with respect to α under the constraints $0 \leq \alpha_i \leq \gamma$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Note, that the dual problem is solved in the object space which is n-dimensional and therefore the SVM can be solved efficiently even for $p \gg n$ problems with a computational time $O(n^3)$. In order to uniquely characterize the solution to the primal and the dual problems the Karush-Kuhn-Tucker (KKT) conditions must be satisfied; Karush (1939). This means that if we have the solution to either the primal or dual problem under these conditions we can also use the conditions to obtain the solution to the dual or primal problem, respectively. The KKT conditions are the equations (3.26)-(3.28) and

the following three conditions

$$\begin{aligned} \alpha_i[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)] &= 0 && (3.30)\\ \mu_i\xi_i &= 0 && (3.31)\\ y_i(x_i^T\beta+\beta_0)-(1-\xi_i) &\geq 0 \quad , && (3.32) \end{aligned}$$

for $i = 1, ..., n$.

The dual problem L_D in (3.29) is defined entirely by inner products of the input space x. So, if we transform the feature space to $h(x)$ then L_D will have the following form

$$L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j h(x_i)^T h(x_j) \tag{3.33}$$

which again involves the transformed features space $h(x)$ via inner products. This can be exploited such that we do not need to explicitly specify the transformation $h(x)$, but only need to specify the kernel function which is defined as the inner product of the transformed feature spaces:

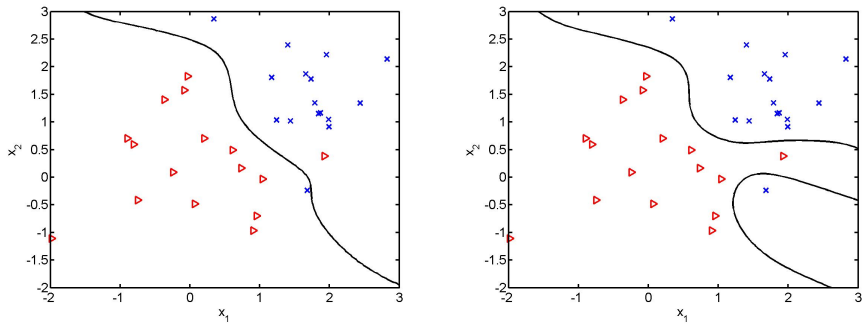
$$K(x_i, x_j) = \left\langle h(x_i)^T h(x_j)\right\rangle \quad . \tag{3.34}$$

This is also called the *kernel trick*. The kernel function should be positive definite such that the Lagrangian dual function L_d is bounded below. The kernel function can be an infinite space without us having to explicitly specify such a space. This is e.g. the case with the radial basis function $K(x_i, x_j) = \exp(-\kappa\|x_i - x_j\|_2^2)$. The kernels give non-linear separations in the original input space x (corresponding to linear separations in $h(x)$). With the use of kernels the regularization weight γ is very important since the risk of overfitting increases when we expand the input space (in an enlarged input space two classes will often be perfectly separable). In figure 3.9 the radial basis kernel was used to separate the data from figure 3.8 for two values of γ.

From (3.26) it is seen that the solution function $\phi(x)$ can be written as:

$$\phi(x) = \sum_{i=1}^{n}\alpha_i y_i K(x_i, x_j) + \beta_0 \quad . \tag{3.35}$$

When $\alpha_i = 0$ the i^{th} point does not contribute to the solution. The remaining points for which $\alpha_i > 0$ are the support vectors. When $0 < \alpha_i < \gamma$ the i^{th} point lies on the margin, and for these points $\xi_i = 0$ and we can use (3.30) to estimate β_0 (a numerically stable solution is obtained by averaging for all the margin points).



(a) $\gamma = 10^{-3}$, 40 support points (b) $\gamma = 10^{3}$, 28 support points

Figure 3.9: An example of a nonlinear SVM with a radial basis function with $\kappa = 0.5$ and for two values of the regularizing parameter γ. It is seen that when we use a low value of γ all points are included in the solution (all points are support points) and we get a less complex function. On the other hand, when we use a higher value of γ fewer of the points are included in the solution, and the solution function becomes more complex.

To sum up we will state the key features of SVMs by using the words of Shawe-Taylor and Cristianini (Shawe-Taylor and Cristianini, 2004): "The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin."

Finally, we note that the SVM has been parallelized for optimal speed in Parallel SVM (PSVM), open source. NIPS 2007, Chang et al.

In chapter 7 (paper D), we add a constraint to the SVM which can hold information about e.g. pairing of data.

3.6 Regularization of the parameter estimates

One way of addressing large p, small n problems is by regularization of the parameter estimates. The general loss function for regularizations looks as follows

$$L(X, \beta) = ERR(X, \beta) + \lambda P(\beta) \quad , \tag{3.36}$$

where $ERR(X, \beta)$ is the estimate of our model error with parameters β such as the MSE when we use a regression model. $P(\beta)$ is a penalization term of the model parameters with weight λ. The regularization can be written as an

equivalent constrained problem

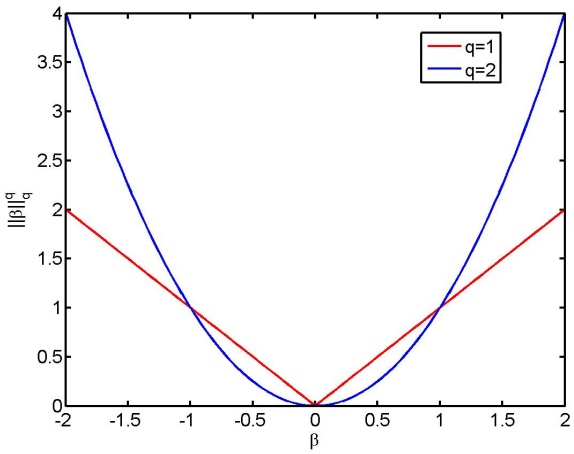
$$L(X,\beta) = ERR(X,\beta) \quad \text{s.t } P(\beta) \leq t \quad . \tag{3.37}$$

In general the penalization term punishes the model complexity. Two well known regularization models of this kind are the ridge and the lasso models.

The ridge model (Hoerl and Kennard, 1970) penalizes the ℓ_2-norm of the parameter estimates β, see figure 3.10. The loss function for ridge regularization is

$$L_2(X,\beta) = ERR(X,\beta) + \lambda_2 \|\beta\|_2^2 \quad , \tag{3.38}$$

where the ℓ_2-norm is defined as $\|\beta\|_2 = \sum_{j=1}^{p} \beta_j^2$ λ_2 is the penalization weight on the parameter prior. This means that all parameter estimates are shrunken towards zero and in particular extreme parameter values are punished under this norm.



(a) Norms in 1D

Figure 3.10: Illustration of the ℓ_1- and ℓ_2-norms.

The lasso (least angle selection and shrinkage operator) (Tibshirani, 1996; Efron et al., 2004) penalizes the ℓ_1-norm of the model errors, see figure 3.10. The loss function for lasso regularization is

$$L_1(X,\beta) = ERR(\beta) + \lambda_1 \|\beta\|_1 \quad , \tag{3.39}$$

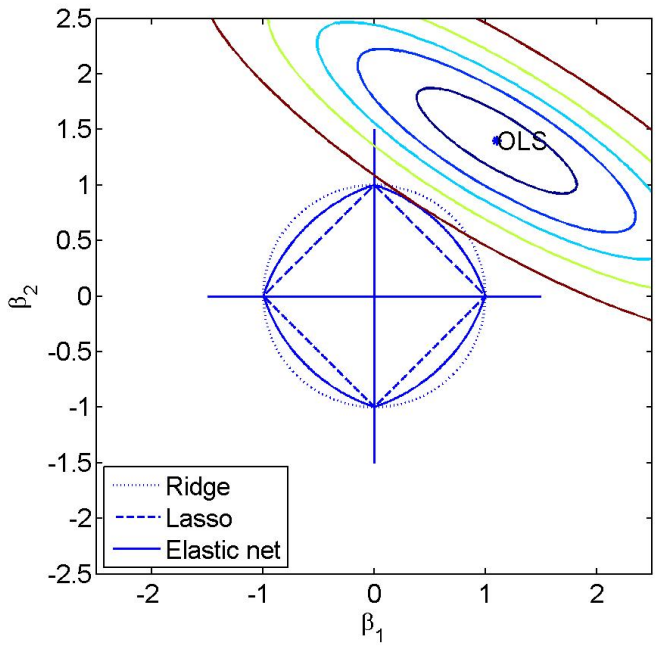
where the ℓ_1-norm is defined as $\|\beta\|_1 = \sum_{j=1}^{p} \text{abs}(\beta_j)$ and λ_1 is the penalization weight on the parameter prior. Under this norm there is a large possibility of setting one or more of the parameter estimates to zero, the non-zero parameter

estimates are shrunken moderately and the ℓ_1-norm does not punish outliers as hard as the ℓ_2-norm does.

Recently, the elastic net which adds both the ridge and the lasso constraints to the OLS setting was proposed (Zou and Hastie, 2005). The loss function for the elastic net combination is

$$L_{en}(X, \beta) = ERR(\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \quad . \tag{3.40}$$

Regularizations of this kind correspond to/are a prior on the parameter estimates. This means that we make an assumption about the distribution of the parameter estimates as illustrated in 3.11. The figure illustrates the OLS solution and the contour curves of the MSE function of β. Where the contour curves intersect with the priors on the model parameters we find the solution to the ridge, lasso, and the elastic net (which is a model where both the ℓ_1- and ℓ_2-norms are added as regularizing terms), respectively.



(a) Constrained norms in 2D and OLS solution

Figure 3.11: Illustration of the ℓ_1- and ℓ_2-norms.

3.6.1 Advantages and disadvantages of using the ℓ_1-norm prior

The biggest advantage of the lasso is that it performs a feature selection, and thus reduces the dimensions. The disadvantage is that the lasso may be too sparse in many settings Bro (2009). Donoho thus also showed in Donoho (2006) that in large p, small n problems the lasso solution most often is the sparsest solution. However, this does often make the lasso favorable over greedy algorithms such as forward selection. An additional disadvantage of the lasso is that at most n features can be included in the solution before it saturates (the model overfits).

With regards to solving the lasso problem there exists fast algorithms such as basis pursuit denoising (BPD; Shaobing and Donoho (1994)) and least angle regression selection (LARS; Efron et al. (2004); Hesterberg et al. (2008)) despite the fact that the ℓ_1-norm is non-differentiable.

3.6.2 Advantages and disadvantages of using the ℓ_2-norm prior

The advantages of the ℓ_2-norm prior are that it shrinks the parameter estimators towards zero and thereby ensures more generalizable solutions as the effective dimensions are reduced, and it is a quadratic constraint and therefore quadratic optimization problems remain convex under this constraint. The disadvantage is that it does not perform feature selection and thus redundant and irrelevant variables contribute to the solution.

The ridge problem is easy to solve as it is a linear operator due to the differentiability of the ℓ_2-norm. It can be solved in the n-dimensional space which is often a considerable reduction in large p, small n problems.

3.6.3 Advantages of using both the ℓ_1 and ℓ_2-norm priors

The advantages of including both the ℓ_1 and ℓ_2-norm priors are that they can perform feature selection and shrinkage of parameters and thus dimension reduction and generalizable solutions are obtained. Jointly they also give a grouping effect of variables which means that highly correlated features are given similar coefficients. The author has good experience with the elastic net and the

grouping effect when correlations exist in the feature set as for example with spatial and spectral correlations in the examples concerning multi-spectral images. Thus together the regularizations overcome most of the disadvantages of using only the ℓ_1-norm or only the ℓ_2-norm. The major disadvantage is that there are two parameters to tune. In chapter 5 we give further details on the elastic net, and make comparisons of the elastic net with various other dimension reduction techniques.

3.7 Model selection

Cross-validation (CV) is a method which can be used for comparison of models, for example to compare models with varying numbers of dimensions, or models with different underlying assumptions on distributions etc. Cross-validation gives an estimate of the expected prediction error. For K-fold cross-validation the training data is split randomly into K parts and in turn each part is left out and predicted by fitting a model on the remaining $K-1$ parts of the data. The cross-validation error is then the average prediction error on the K parts left out. This setting is illustrated in figure 3.12. When we have a particularly small number of samples we can use *leave-one-out* CV where $K = n$ meaning one sample is left out in turn and a total of n models is built. CV is a frequentist approach as it uses observed data to assess the model. It is powerful for estimating the prediction error and thereby comparing various methods.

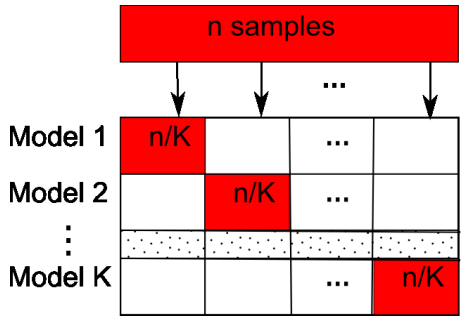


Figure 3.12: Illustration of the cross-validation technique. The red parts of the data set are the parts we leave out in each model. A model is then built on the remaining $K-1$ parts and the left out red data is predicted.

The major drawback of CV is that it can be computationally costly to fit K models as K increases. The computational load further increases when there

are several parameters which must be set for a given model; e.g. the number of dimensions, the weights for the penalization terms etc.

In the present thesis we in general use cross-validation for validation of our models with regards to not overfitting training data, i.e. we use it to tune the model parameters. We then use a separate test set to estimate the prediction error and compare models of various kinds.

Cross-validation can be seen as sampling without resubstitution, another approach is to use sampling with resubstitution this is also known as *boot-strapping* and is in particular useful when samples are very scarce.

An alternative to the data driven model selection techniques is to use Bayesian methods to set the parameters. In Bayesian modelling overfitting is avoided by integrating or summing over all sets of possible parameters rather than making a point estimate of their values based on the observed data (Bishop, 2006, section 3.4). The integration often requires that Monte Carlo sampling is performed and can therefore often be computationally costly Bishop (2006). One example where the degree of sparseness of a model is estimated with expectation maximization (EM) in a Bayesian setting is given in Figeuiredo and Jain (2001). The cost of this way of estimating the parameters is $O(n^3)$.

Chapter 4

Paper A - Multiplicative updates for the LASSO

The paper included in A introduces an algorithm which uses multiplicative updates (MU) for solving the least absolute shrinkage selection operator (lasso; section 3.6) problems. MU are traditionally used for non-negativity constrained problems, but here it is illustrated how they can be used for unconstrained problems as well.

It gives an intuitive and easy to implement algorithm for lasso. As lasso is one of the basic methods used for feature selection, this paper is placed first in the methodical part of this thesis. The algorithm can easily be extended to other types of cost functions, e.g. include more or other types of priors on the parameter estimates.

MU is proven to monotonically decrease the cost function. The idea behind MU is to use the relation between the negative and the positive part of the gradient of the cost function to update the parameter estimates. For this purpose the problem is expanded with the negative and positive parts of the parameter estimates. This is done to avoid singularities of the derivative of the ℓ_1-norm. The structure of the expanded lasso problem is well exploited with MU compared to traditional quadratic programming (QP) solvers and is thus computationally faster.

Results are given for MU, a standard QP solver, and basis pursuit denoising (BPD) algorithms of the lasso problem on three bioinformatics data sets. Two of which are large p, small n problems with microarrays from different groups of cancer patients. MU outperformed the QP solver, but was not as fast as BPD on large problems, and also converged slower than BPD for small values of the regularization parameter (when there were many non-zero parameter estimates).

Chapter 5

Paper B - A comparison of dimension reduction methods with applications to multi-spectral images of sand used in concrete

This paper, included in B, describes the elastic net which combines a prior of the ℓ_2- and ℓ_1-norm of the parameter estimates with regression analysis (section 3.6). Furthermore, a comparison is performed of the elastic net against forward selection, principal component regression, and a combination of a genetic algorithm for feature selection and a partial least squares modelling.

With regards to the present thesis, this paper serves as an introduction to the least angle regression selection (LARS) technique which is a less greedy feature selection algorithm than e.g. the traditional forward selection algorithm. With simple modifications it can furthermore evaluate lasso and forward stagewise solutions. It can at the computational cost of a single ordinary least squares fit give solutions for the entire parameter path (for all sets of parameter estimates). It is also shown how LARS can be used to efficiently compute elastic net solu-

tions. Likewise, it serves as an introduction to the elastic net and compares it with a range of classical and one more sophisticated method as mentioned in the previous paragraph.

As an example the moisture contents of sand samples were predicted from five different types of sand used to make concrete. This was facilitated using multi-spectral images such that the prediction eventually can be performed non-invasively and in the construction line.

The conclusions were that the elastic net and genetic algorithm - partial least squares in general gave the best generalizations in form of the lowest predicted error rates (leave-one-out cross-validation rates) and overfitted data less. Finally, the elastic net gave sparser solutions (fewer non-zero parameter estimates) than the combination of genetic algorithm feature selection and partial least squares. The standard deviations of the moisture content estimates were typically around 0.5% moisture content from samples with moisture contents varying from 1.25%-8.75% moisture content.

Chapter 6

Paper C - Sparse Discriminant Analysis

This paper, included in C, extends both linear and mixture discriminant analysis to large p, small n problems (section 3.3 and 3.4). The linear discriminant analysis and mixture discriminant analysis have previously been rewritten to the equivalent regression type problems called optimal scoring. In this setting the ridge and lasso penalties are added to obtain sparse, general solutions to the classification problems. The methods provide low dimensional views of data which each are based on only a subset of features.

The methods were compared to penalized discriminant analysis, shrunken centroids regularized discriminant analysis, sparse partial least squares, and forward selection based on Wilk's Λ combined with linear discriminant analysis. To evaluate the performance of the methods the following examples were used: A microarray data set comparing gene expressions for six different subtypes of cancer, classification of *Penicillium* fungi to the species level based on multispectral images, classification of three genera of microbiological fungi based on 1D NMR spectra, and classification of three fish species based on RGB images.

The sparse discriminant analysis and sparse mixture discriminant analysis performs comparably to shrunken centroids regularized discriminant analysis. Also, it outperformed the other methods with respect to classification rates on the

given examples. Furthermore, the proposed algorithms gave sparser solutions than the methods for comparison, i.e. fewer of the original variables were included in the solutions.

Algorithms have been developed in both R and Matlab and these are available through CRAN and my website.

Chapter 7

Paper D - Classification of paired ear canal impressions in high dimensions - Data driven constraints for the Support Vector Machine

Paper D adds a prior for paired data to the support vector machine (SVM; section 3.5). The problem at hand is to classify ear canal impressions into whether they were taken with open or closed mouth. This is an advantage for the hearing aid manufacturer as there amongst doctors does not exist a standardized way of obtaining the ear canal impressions for hearing aid patients.

A difference vector was created going from the open mouth to the closed mouth setting for each paired observations. This gives a description of the change in shape of the ear canal when an individual opens the mouth. The a priori information of the paired observations is added to the SVM as a penalization of non-orthogonality between the difference vectors and the normal vector to the separating hyperplane.

The orthogonality constraint gives similar or slightly improved classification rates compared to the SVM for the ear canal example and on top of that reduces the variance of the solution by 50%.

The optimization problem with the constraint added to the SVM is derived as a general constraint and thus the theory can be utilized for other kinds of constraints as well. Such constraints could for example be taking into account correlations between features as for example in fused lasso.

Part II

Applications

This part includes three papers and an unpublished report with various applications of the methods described in the previous part of this thesis. They should not necessarily be read in chronological order, but can be read separately. I can give no better advice than to read only the papers/chapters for which the application/example is of interest to the reader.

Chapter 8

Paper E - Multi-spectral recordings and analysis of psoriasis lesions

Paper E uses the elastic net to predict the psoriasis area and severity index (PASI) from features extracted from multispectral images. The PASI scores were first evaluated for each lesion by a trained doctor, then multi-spectral images were acquired of the lesions and the scores were calculated based on these.

Using multi-spectral images for the analysis of the psoriasis lesions makes the evaluations objective and standardized. The PASI evaluations are widely used to evaluate psoriasis patients and propose suitable treatments.

A simple threshold was used to segment the lesions from each image and features were calculated which are first order statistics of the reflectance values, and for the pairwise ratios and differences between reflectance values of the spectral bands.

Two models were made: One for the eurythema and one for the infiltration of the psoriasis lesions. The elastic net was used to regress the PASI scores which range from 0 to 4 (more on the elastic net in chapter 5). Only two features

were selected for each model and the standard deviations were 0.5 and 0.6, respectively.

The paper is, in the version that is included here, an extended version of that in the original proceedings. The most important amendment is that results on the spatial coherence between the percentile features have been included. These illustrate how the summary statistics used as features to represent the region of interest in a multi spectral image can be seen as more robust features than e.g. spectral reflectances per pixel. Such a representation exploits the high correlation between pixels and thus one of the blessings in such high dimensional problems; see section 2.2. Furthermore, it is illustrated how the spatial information is not completely lost when the summary statistics are used as features.

The paper illustrates how multispectral images can be used as an objective means for quantifying the severity of psoriasis lesions.

Chapter 9

Paper F - Individual discriminative face recognition models based on subsets of features

Paper F uses the elastic net to find subsets of features which are suitable for face recognition (more on the elastic net is found in chapter 5).

The features consist of both landmark coordinates as well as red, green, and blue color intensities of matched faces from the XM2VTS database. One-against-all classification models were built on four training images of each subject and then two images were used for validation and parameter tuning, and finally another two images were used for testing the methods.

The elastic net was compared to various other face recognition methods. The elastic net compares to methods such as Eigenfaces and Procrustes on the colorimetric and geometric features, respectively. However, it does not give as good classification results as Fisherfaces.

It was illustrated how the elastic net can be utilized to find small subsets of features which are relevant for the task at hand and still obtain good performance

although not comparable to state-of-the art face recognition methods.

Chapter 10

Paper G - Temporal reflectance changes in vegetables

Paper G makes use of the elastic net to identify subsets of features which describe color changes in celeriac and carrots over time. Such changes in appearance are used for quality assessments of foods in general. Statistical tests are carried out to test if there is a significant difference in the reflectance values of the identified feature subsets over time.

The vegetables were stored in a refrigerator and a handful of each kind were taken out at day 2, 4, 8, 10, 12, and 14 of the experiment and digitized using a multispectral camera (called VideometerLab). A watershed algorithm was utilized to segment each piece of carrot and celeriac in the images. First order statistics (1, 3, 5, 10, 20, 40, 50, 60, 80, 90, 95, 97 and 99th percentiles) of the reflectances and the pairwise ratios between reflectances were extracted as features. The problems were of a total of 3249 features and around 400 observations for both carrots and celeriac. Regression models were built using the elastic net technique; more on the elastic net in chapter 5. Based on the selected features pairwise two-sided Bonferroni corrected t-tests were carried out to check if the differences in reflectances over the days were significant. For the carrots, there was a significant change from day 2 to day 4, but no significant

change after that. For the celeriac there were significant changes up till day 12. Furthermore, the selected subsets of features indicate that the changes which appeared were most likely caused by oxidation.

The paper illustrates how multispectral imaging can be used for quality assessments of foods.

Chapter 11

Classification and identification of Aspergillus fungi based on multi-spectral images

11.1 Abstract

This report illustrates the difficulties of classification of the two *Aspergillus* species *niger* and *tubingensis*. In particular, it illustrates the difficulties of identification of black *Aspergillus* strains based on morphological information. Multi-spectral imaging have been used to obtain an objective method for classification of the two species. The multi-spectral images acquire high resolution images in 18 spectral bands; 10 in the visual range (400-700nm), and 8 in NIR (700-1000nm). The spectra in the visual range carry morphological information whereas the NIR spectra additionally provide some chemical information. The results give a classification rate of 88% for known strains and also illustrates that identification of new black *Aspergillus* strains are practically impossible without a chemical or molecular analysis.

11.2 Introduction

Aspergillus niger is used for production of both citric acid and a variety of enzymes used in food industries. It has previous, with a few restrictions, been categorized as a safe production organism Schuster et al. (2002). These facts make it one of the most important microorganisms used in biotechnology. *Aspergillus tubingensis* can also be used for citric acid production, but is less known than *A. niger*. The two species *A. niger* and *A. tubingensis* are considered distinguishable with molecular data only, see e.g. Schuster et al. (2002), Gonzalez-Salgado et al. (2005). They belong to the section *Nigri* of the *Aspergillus* genus, are both dark brown to black in conidium color, and their genomes have high similarity Juhasz et al. (2008).

As chemical and molecular analyses are expensive and time consuming it is of interest to perform simpler analyses to classify the species, such as traditional morphological analysis. However, as the two species are so closely related morphological classification has been disregarded as a good mean of classification and identification of, in particular black strains of the *A. niger* and *A. tubingensis* species. Here, we investigate the use of an objective method based on multi-spectral images for fast and easy assessment to a classification. The images are based mainly on visual spectra which are related to the morphology of the strains, but also includes some near infra red (NIR) spectra providing knowledge of their chemistry as well. The objective analysis adopted here is based on multi-spectral images and is similar to the analysis performed on species of the *Penicillium* genus in Clemmensen et al. (2007).

11.3 Data

The present experiment includes 24 *A. niger* strains and 8 *A. tubingensis* strains. There are 2 sets of replica (in one case 4) which were inoculated by two individuals on two growth media: YES and CYA. The samples were incubated for 7 days at 25°C. The 2 sets of replica are split into 33 training samples and 33 test samples (one set of replica for each). Table 11.1 lists the strains used in the present experiment.

Table 11.1: List of the strains used in the presented experiment.

Strain	Species	Replica
IBT 26387	*A. niger*	2
IBT 24634	*A. niger*	2
IBT 26392	*A. niger*	2
IBT 20959	*A. niger*	2
IBT 23191	*A. niger*	2
IBT 27876	*A. niger*	2
IBT 24631	*A. niger*	2
IBT 19348	*A. niger*	2
IBT 18741	*A. niger*	2
IBT 26391	*A. niger*	2
IBT 25744	*A. niger*	2
IBT 12710	*A. niger*	2
IBT 23432	*A. niger*	2
IBT 25753	*A. niger*	2
IBT 26774	*A. niger*	2
IBT 28086	*A. niger*	2
IBT 25752	*A. niger*	2
IBT 13099	*A. niger*	2
IBT 19085	*A. niger*	2
IBT 22447	*A. niger*	2
IBT 20963	*A. niger*	2
IBT 20381	*A. niger*	2
IBT 23721	*A. niger*	2
IBT 26773	*A. niger*	2
Strain	Species	Replica
IBT 4356	*A. tubingensis*	2
IBT 3567	*A. tubingensis*	2
IBT 26390	*A. tubingensis*	2
IBT 23420	*A. tubingensis*	2
IBT 25339	*A. tubingensis*	2
IBT 28110	*A. tubingensis*	2
IBT 28115	*A. tubingensis*	2
IBT 3253	*A. tubingensis*	2
IBT 28115	*A. tubingensis*	4

11.3.1 Multi-spectral imaging

The multi-spectral images were acquired with VideometerLab. The 18 spectral bands were acquired at the wavelengths: 430, 450, 470, 505, 565, 590, 630, 645, 660, 700, 850, 870, 890, 910, 920, 940, 950, and 970nm. The images are in the resolution: 960×1280 pixels and are taken with a radiometric resolution of 8 bits/pixel. The light setting in VideometerLab is controlled and reproducible over time, see Gomez et al. (2007).

11.3.2 Masking out the fungal colonies

The region of interest (ROI) in the images are the fungal colonies. WE therefore want to mask the fungal colonies out from the background, the petri dish, and the growth medium. The

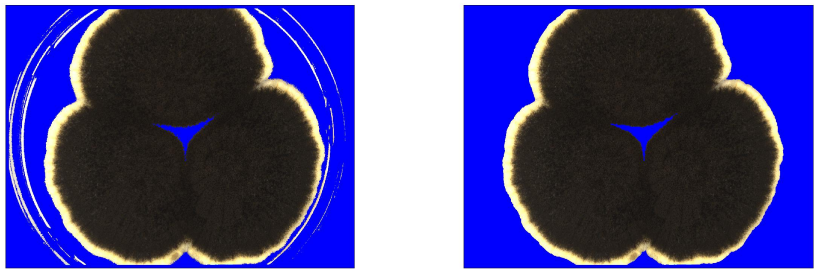
Find ROI based on spectral information in each pixel, see Figure 11.1:

- Separate pixels with spectral information of interest from background, medium and petri dish: $ROI_1 = (b_{18} - b_1) > 30 \vee (b_{16} - b_2) > 30 \vee b_1 > 50$. b stands for spectral band.
- Some reflections of light in the petri dish have not been removed by the spectral separation. Remove this by performing a morphological image opening (erosion + dilation) on ROI_1 w. a disk structure element (strel) of size 10: $ROI_1 = imopen(ROI_1, strel('disk', 10))$.

Reflections of the fungal colonies into the petri dish have not been removed in ROI_1, see an example in Figure 11.3. These are removed by simple edge detection of the petri dish and by using the fact that the dish is circular, see Figure 11.2:

- The petri dish is located by simple edge detection on band 18. That is, 4 points are located on the edge using the change in reflectance, and a circle is fitted to the 4 points. The center of the circle is denoted: (x_c, y_c).
- A smaller circle (-15 pixels in radius) is created and everything outside this radius (r_2) is masked out: $ROI_2 = \sqrt{(x - x_c)^2 + (y - y_c)^2} < r_2$.

Finally the mask which identifies pixels of fungal colony is created as: $MASK = ROI_1 \wedge ROI_2$.



(a) Before opening (b) After opening

Figure 11.1: Example with light reflections in the petri dish. Sample IBT 23191, replica a, before and after morphological image opening. This figure corresponds to ROI_1.

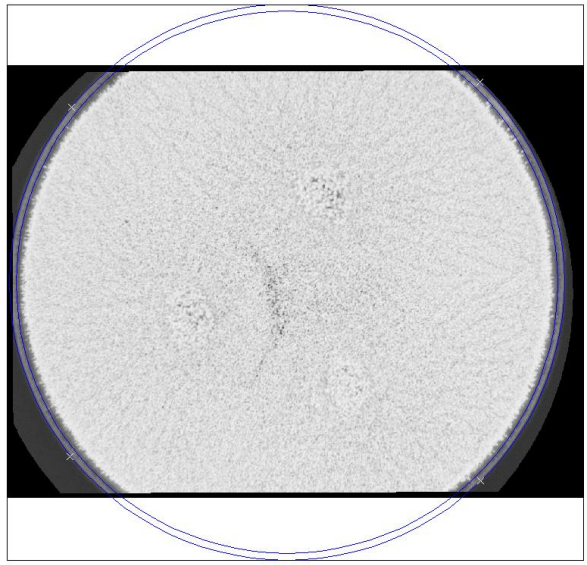
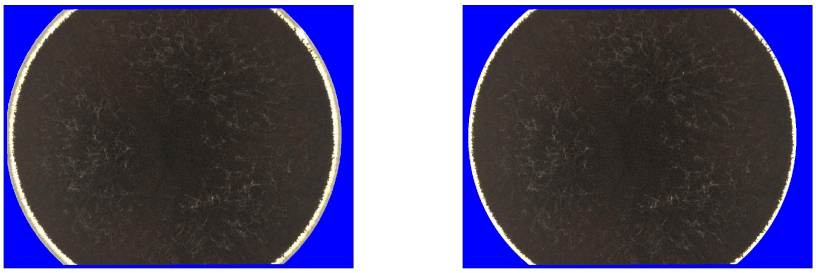


Figure 11.2: Edge detection and circle fit to remove the petri dish, ROI_2. Spectral band 18 of sample IBT 12710, replica a.



(a) ROI_1 (b) $ROI_1 \wedge ROI_2$

Figure 11.3: Example with reflections of fungi in the petri dish. Sample IBT 12710, replica a, after morphological image opening (ROI_1), and after also removing the petri dish by edge detection $ROI_1 \wedge ROI_2$.

11.3.3 Summarizing features of the colonies

There is still a vast amount of pixels present in the region of interest identified by the mask described in the previous section. To summarize the information present in the fungal colonies we use basic, summarizing statistics. The following features were extracted from the pixels of the fungal colonies:

- Standard deviation, mean value, maximum intensity, and the $(1, 5 : 5 : 95, 99)$th percentiles (21 percentiles) of the 18 spectral bands, the 152 pairwise differences between the spectral bands, and of the 152 pairwise multiplications between the spectral bands.

- The $(5 : 5 : 100)$th percentiles of the median, variance, and mean texture images of each of the 18 spectral bands (in total 54 texture images).

The first part of the features are based on spectral information (7776 features/sample) whereas the remaining part is based on textural information (1080 features/sample).

11.4 Methods

11.4.1 Sparse discriminant analysis

Sparse discriminant analysis (SDA) was used to make a classification model Clemmensen et al. (2009a). The method finds sparse and penalized directions which discriminates the classes most. It is an extension of linear discriminant analysis, in particular useful when the number of variables by far exceeds the number of observations. The method performs a linear discriminant analysis (LDA, Fisher (1936)) with variable selection using the ℓ_1-norm (Lasso, Tibshirani (1996)), and a coefficient shrinkage using the ℓ_2-norm (Ridge, Hoerl and Kennard (1970)). In this way a robust classification with an easy interpretation can be reached. The method has two parameters which need to be tuned: The number of active variables, and the weight on the coefficient shrinkage (λ).

11.4.2 Cross-validation

In order to select the two parameters in the classification method cross-validation over the training set is used, leaving the test set independent. A leave-one-out cross-validation is used since the number of observations is small (33), see Duda et al. (2001). One observation is left out a time and a model is build over the rest of the training data, the observation left out is then classified using the model built. The parameters are selected such that the classification error over all the, in turn left out, observations are minimized.

11.4.3 Test for further information

As in Clemmensen et al. (2007) we will use Mahalanobis distance and the *test for additional information*, see e.g. Rencher (2002) for further details on these statistics. The null-hypothesis of the test states that the last q variables do not contribute to a better discrimination. The tests were conducted at a 5% level of significance. As in Clemmensen et al. (2007) the first principal components are used as variables for these tests.

11.4.4 Unmixing of spectra

Mixture models were used to model the spectrum of each pixel in a sample image consisting of horizontal scan lines from all the samples. As the spectral shape of the fungi are unknown an algorithm called *iterated constrained endmembers* (ICE, Berman et al. (2004)) were used to find endmembers in the image. Each spectrum is then estimated using a mixture model with these endmember, as in *the spectral assistant* (TSA, Berman et al. (1999)). The software used here was VoiR, a licensed R-package developed by CSIRO, Australia.

11.5 Results

11.5.1 Classification of strains

The parameters selected with leave-one-out cross-validation on the training set were: 40 non-zero loadings, and $\lambda = 10^{-3}$. Building a model with these parameters on the 33 training samples and then testing with the 33 test samples gives a training classification of 100%, and a test classification of 88%. The confusion matrices for the training and the test sets are illustrated in Table 11.2. The four misclassified observations are replica of the strains: IBT 12710, IBT 20963, IBT 25339, and IBT 3253. Figure 11.4 illustrates the sparse discriminant direction which is the base of the classification model. Furthermore, estimated Gaussian distribution for the two classes are plotted to illustrate the separation. Note, how these distributions change from peaks (having small variance) to more wide bells (having large variance) when the test data is included in the estimation. Furthermore, the means of the distribution shifts closer to each other when the test data is included. The big shift in the distributions show, as well as the classification rates, that the training data is slightly overfitted.

Table 11.2: Confusion matrices for the training and test sets.

Train	T	N	error
T	9	0	0.0000
N	0	24	0.0000

Test	T	N	error
T	7	2	0.2857
N	2	22	0.0909

Selected features are listed in Table 11.3. The majority of the intensity based features (spectral information rather than texture based information) includes NIR spectra. This illustrates that the chemistry rather than the morphology is useful as a mean of classification of the two species.

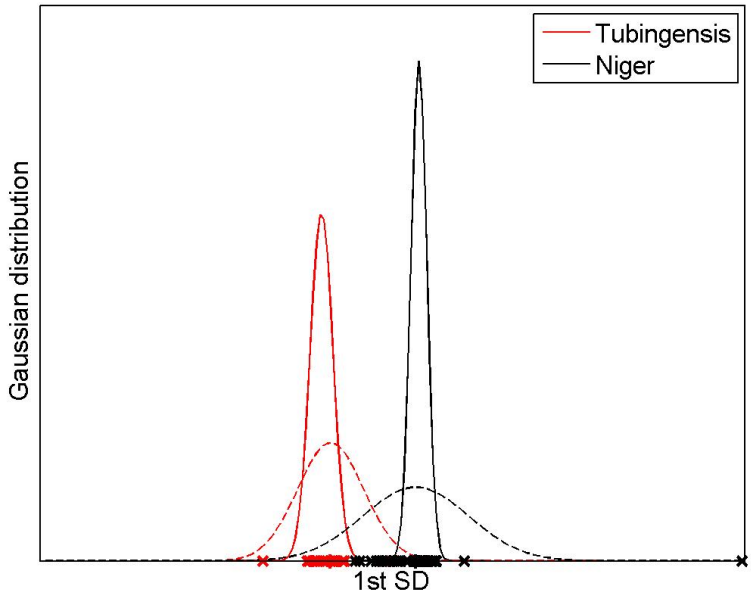


Figure 11.4: Illustration of the data projected onto the sparse discriminant direction. The crosses are the observations, the solid lines are estimated gaussian distributions of the two species for the training data, and the dashed lines are estimated gaussian distributions of the two classes for the training and test data.

Table 11.3: List of the selected features.

Medium	Image	Feature
CYA	band 17	Max intensity
CYA	band 1 - band 2	Max intensity
CYA	band 1 - band 3	99th percentile
CYA	band 6 - band 7	99th percentile
CYA	band 9 - band 10	1st percentile
CYA	band 9 - band 10	5th percentile
CYA	band 10 - band 14	Max intensity
CYA	band 10 * band 17	Max intensity
CYA	band 13 - band 15	Max intensity
CYA	band 13 * band 15	Max intensity
CYA	band 14 - band 17	Standard dev.
CYA	band 15 * band 18	Standard dev.
YES	band 18	Standard dev.
YES	band 18	Max intensity
YES	band 1 - band 2	Max intensity
YES	band 5 - band 7	85th percentile
YES	band 5 - band 7	90th percentile
YES	band 12 * band 18	Max intensity
YES	band 13 - band 14	Max intensity
YES	band 13 * band 15	Max intensity
YES	band 14 - band 17	Standard dev.
YES	band 17 * band 18	90th percentile
CYA	var texture of band 8	10th percentile
CYA	var texture of band 8	15th percentile
CYA	var texture of band 8	20th percentile
CYA	var texture of band 9	20th percentile
CYA	var texture of band 9	25th percentile
CYA	median texture of band 10	100th percentile
CYA	var texture of band 11	65th percentile
CYA	var texture of band 11	70th percentile
CYA	var texture of band 11	80th percentile
CYA	var texture of band 11	85th percentile
CYA	var texture of band 11	90th percentile
CYA	var texture of band 11	95th percentile
CYA	var texture of band 11	100th percentile
CYA	median texture of band 18	5th percentile
CYA	mean texture of band 18	5th percentile
YES	var texture of band 15	20th percentile
YES	var texture of band 15	25th percentile
YES	var texture of band 15	30th percentile

A method called *random forests* Breiman (2001), which builds a number of random trees and then let them vote for the class, was also explored to test if a non-linear classification model could improve the results. However, this method only made the overfitting worse, even when feature selection was used.

11.5.2 Identification of strains

The following results are on a completely independent test set, i.e. no replica of the strains were included in training the model. The results are illustrated in Table 11.4. None of them are correctly identified using the morphological model. It appears that the colors are "unspecific", i.e. that the colors cannot be specified for the two species, but only for a number of strains of the two species. This illustrates why it has been, and still is, so difficult to identify the black *Aspergillus* strains based on traditional morphological methods.

Table 11.4: Classification of Brazilian species based on the same model as above results. Both replica are classified to the same species for all strains. The "true" species is the identification based on chemical analyses.

Strain	"True" species	Estimated species
BRUZ 2	*A. aculeatus*	*A. niger*
BRUZ 4	*A. niger*	*A. tubingensis*
BRUZ 5	*A. niger*	*A. tubingensis*
BRUZ 6	*A. niger*	*A. tubingensis*
BRUZ 7	*A. tubingensis*	*A. niger*
BRUZ 8	*A. niger*	*A. tubingensis*

11.5.3 Test for additional information

The first 7 principal components (PCs) are used as variables to represent the four feature sets: spectral features (spct) and texture based features (txt) and the YES (Y) medium and the CYA (C) medium, respectively. The 7 PCs represent approximately 97% of the variance of the respective feature set in each case. Table 11.5 summarizes Mahalanobi's squared distances and the corresponding p-value of Hotelling's two sample t-test (T^2) stating that the two groups have the same mean.

We note that the spectral features from CYA seems to be the only feature set which with statistical significance can separate the species on its own. Further-

Table 11.5: Mahalanobi's squared distances for the four feature sets. In parentheses the p-values for Hotelling's T^2 test.

Data set	Yspct	Cspct	Ytxt	Ctxt	Ytxt+Ctxt
Yspct	1.0738	1.5807	1.4979	1.5150	3.2597
	(0.10061)	(0.01324)	(0.01890)	(0.01756)	(0.00003)
Cspct		1.5706	1.7259	2.3456	3.6365
		(0.01853)	(0.00709)	(0.00051)	(0.00001)
Ytxt			0.9722	1.9506	-
			(0.14035)	(0.00271)	
Ctxt				0.3998	-
				(0.69022)	
Yspct+Cspct	-	-	2.0653	3.0723	4.3989
			(0.00315)	(0.00006)	(0.00001)

more, as soon as we combine the information from two of the feature sets there is a statistical significant difference between the species.

Now, we will test if the PCs for each of the feature sets provide additional information to the separation of the species. We start with all data sets and remove one feature set at a time. Table summarizes the p-values for the test that the 7 PCs representing the test feature set does not contribute to the discrimination. It is interest to see that although the texture features from

Table 11.6: P-values for Rao's test for additional information.

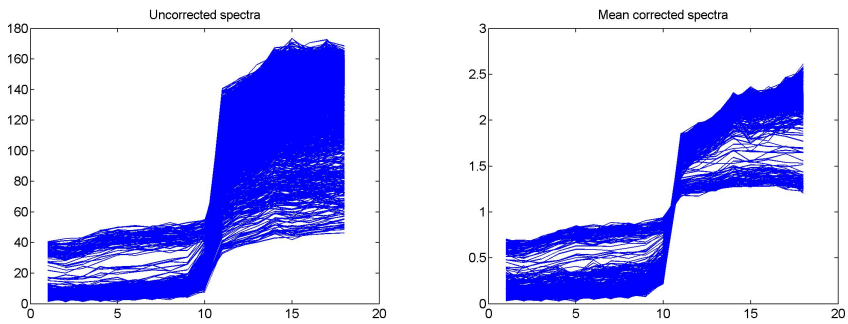
Base feature sets	Test feature set			
	Yspct	Cspct	Ytxt	Ctxt
All	0.85099	0.64607	0.53721	0.12643
Cspct + Ytxt + Ctxt	-	0.18158	0.37402	0.11242
Yspct + Ytxt + Ctxt	0.32841	-	0.13064	0.13580
Yspct + Cspct + Ctxt	0.73262	0.18878	-	0.21691
Yspct + Cspct + Ytxt	0.94151	0.79031	0.85423	-
Yspct + Ctxt	-	-	0.04390	0.25060
Cspct + Ctxt	-	0.02073	-	0.54109
Cspct + Ytxt	-	0.49545	0.98960	-
Yspct + Ctxt	0.18356	-	-	0.80602
Yspct + Ytxt	0.71663	-	0.82156	-
Yspct + Cspct	1.00000	0.74372	-	-

CYA on their own give the poorest separation between the classes they seem to add additional information to the separation in combination with any of the other feature sets. Furthermore, there is no statistical significance, a part from that the texture features from CYA cannot stand alone, which justifies adding

extra features. However, results of Hotelling's T^2 tests in Table 11.5 justifies combining at least two feature sets unless the spectral features from CYA are utilized.

11.6 Spectral analyses

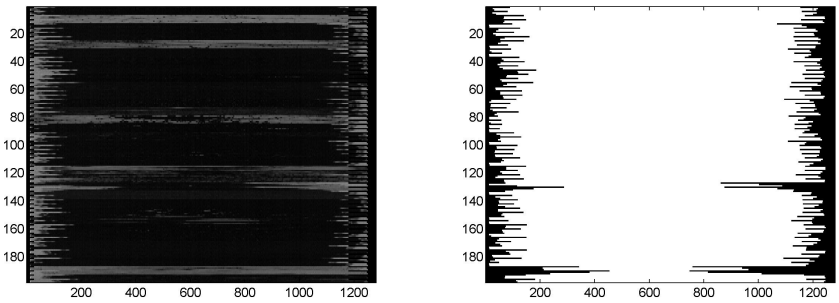
In the following we have taken 3 horizontal scan lines from each image, the lines at pixel: 200, 400, and 600. This gives a total of $3 \times 66 = 198$ rows, still with 1280 columns. This sample image is then analyzed using mixture models as in *the spectral assistant* (TSA, Berman et al. (1999)) in Voir from CSIRO. The Iterated Constraint Endmembers (ICE, Berman et al. (2004)) is applied to find spectral endmembers in the sample image, and proportion maps are constructed using these endmembers. The idea is to see whether one or more endmembers give different proportions for the two species. All spectra were mean corrected, i.e. divided by their mean over all the spectra. The mean correction reduces the variance of the spectra within each sample, see Figure 11.5.



(a) Uncorrected (b) Mean corrected

Figure 11.5: Illustration of the spectra in the three horizontal scan lines of a sample. Sample IBT 24634, replica a, on YES medium.

Figure 11.6 illustrates the sample image with the 198 scan lines and 1280 columns plus the corresponding mask. The last 12 samples are of the *A. tubingensis* species, i.e. the last 36 lines in the sample image should give a distinct pattern if we should be able to tell a difference between the species based on their spectral appearance.



(a) Sample image (b) Corresponding mask

Figure 11.6: The first spectral band of the mean corrected sample image and the corresponding mask.

11.6.1 CYA

A maximum noise fraction decomposition was performed to get rid off noisy bands, the signal to noise ratios are listed in Table 11.7. Based on the first 10

Table 11.7: Signal to noise ratios (SNR) for the 18 MNF bands.

Band no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
SNR	307.9	25.3	16.5	4.1	2.3	1.7	1.6	1.6	1.5	1.5	1.4	1.4	1.4	1.4	1.4	1.3	1.3	0.4

NMF bands the endmembers are identified using ICE. The highest proportions are listed in Table 11.8 for 4 to 8 endmembers. As a rule of thumb the proportions should be higher than 0.6 before an extra endmember is considered. Figure 11.7 illustrates the 7 chosen endmembers and the corresponding purest

Table 11.8: Purest proportion of end members.

No. of ems	em1	em2	em3	em4	em5	em6	em7	em8
4	1	1.0000000	0.8663637	0.8600032	0.0000000	0.0000000	0.0000000	0.0000000
5	1	0.9990102	0.9314493	0.7921452	0.7442423	0.0000000	0.0000000	0.0000000
6	1	0.9562015	0.8467426	0.7622206	0.7063294	0.6965150	0.0000000	0.0000000
7	1	0.9441224	0.7740742	0.7612149	0.7351257	0.6672160	0.6559648	0.0000000
8	1	0.9315595	0.7648188	0.7458720	0.7062171	0.6382836	0.6243022	0.4857641

spectra in the sample image. Figure 11.8 illustrates the proportions maps of the sample image for the 7 endmembers, and the corresponding residual map. There are no obvious patterns in the proportion maps which reveal a difference between the species.

Using linear discriminant analysis for classification on the 211,254 pixels of inter-

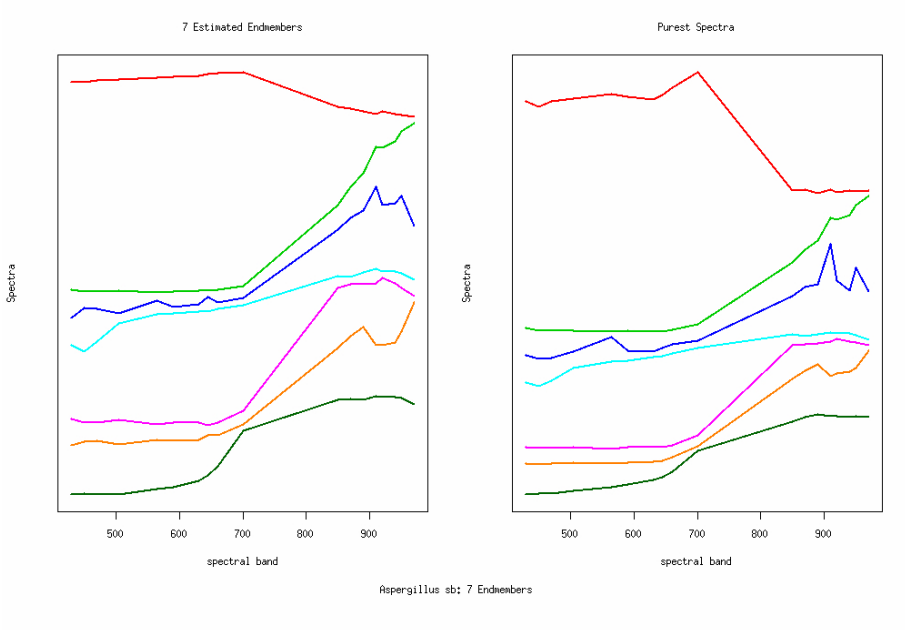


Figure 11.7: The 7 end members and their purest spectra in the samples.

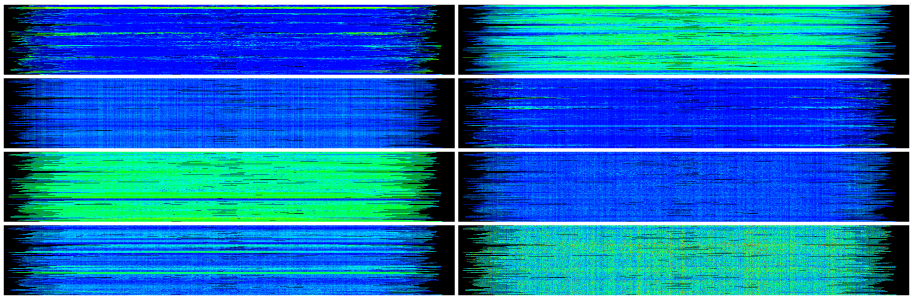


Figure 11.8: The proportions of the 7 end members and the residual map.

est (observations) and 7 proportion measures (features) gave a 73% classification rate using all samples (training plus test), and the following confusion matrix of error rates:

Train	T	N	error
T	190	57314	0.9967
N	65	153685	0.0004

The results underline that the proportions are not suitable for classification between *A. tubingensis* and *A. niger*.

11.6.2 YES

The spectral analyses are repeated on for the strains inoculated at the YES medium. Table 11.9 lists the signal to noise ratios, Table 11.10 summarizes the maimum proportions of the endmembers, and Figure 11.9 and 11.10 illustrate the 6 chosen endmembers, their purest spectra, the proportions maps, and the corresponding residual map. There are no obvious patterns in the proportion maps which reveal a difference between the species. From the proportion maps we see that the fungi mainly consist of two spectral appearances (endmember 2 and 5 for CYA and 3 and 5 for YES), one where the intensities increase in the last NIR bands, and one where the intensities drop. Finally, the results using

Table 11.9: Signal to noise ratios (SNR) for the 18 MNF bands.

Band no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
SNR	1137.9	27.8	16.2	7.2	2.8	1.9	1.7	1.7	1.6	1.6	1.5	1.5	1.4	1.4	1.4	1.3	1.3	0.4

Table 11.10: Purest proportion of end members.

No. of ems	em1	em2	em3	em4	em5	em6	em7	em8
4	1	0.9732812	0.9571946	0.6676608	0.0000000	0.0000000	0.0000000	0.0000000
5	1	0.9425179	0.9381710	0.6949816	0.6859283	0.0000000	0.0000000	0.0000000
6	1	0.9570542	0.9255472	0.7731556	0.6656671	0.6376010	0.0000000	0.0000000
7	1	1.0000000	0.8895716	0.8318132	0.6344779	0.6087830	0.5670732	0.0000000
8	1	0.9926336	0.8705209	0.8299693	0.6259893	0.6040144	0.5319467	0.4412426

LDA on the proportion measures were:

Train	T	N	error
T	3083	54191	0.9462
N	1517	159120	0.0094

With a total classification rate of 74%. Again, this underlines the fact that the proportions are not useful for classification between *A. tubingensis* and *A. niger*.

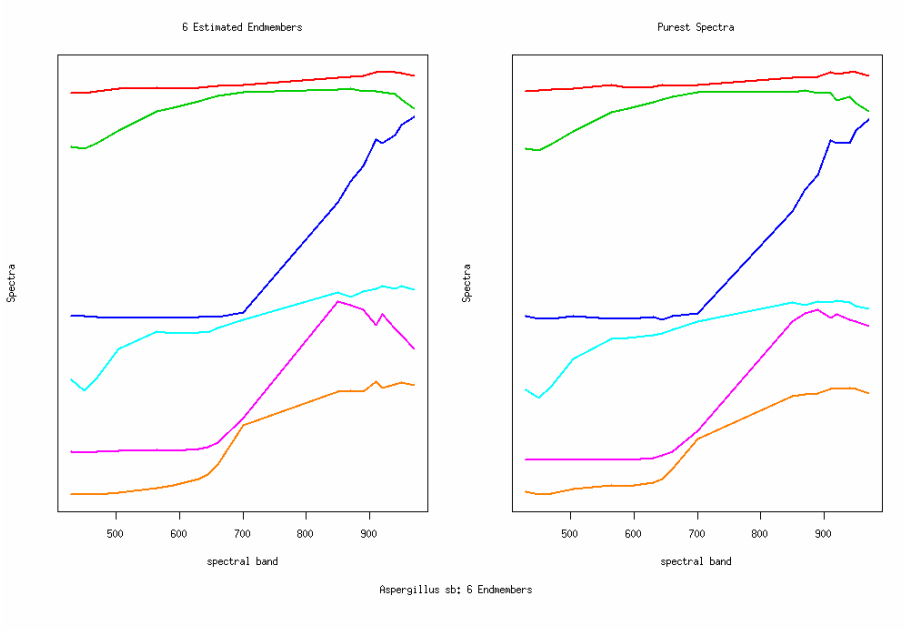


Figure 11.9: The 6 end members and their purest spectra in the samples.

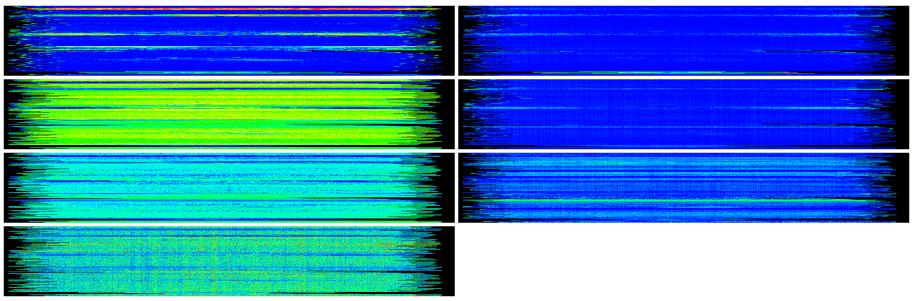


Figure 11.10: The proportions of the 6 end members and the residual map.

11.7 Conclusion

The objective analyses of *A. niger* and *A. tubingensis* illustrated the difficulties in separating the two species by morphology. Although classification of known species is possible to a certain extent (here 88%), there is no basis for giving a general description for classification between them, only one specific for the strains included in training of a model. Additionally, we see that identification is only possible with the use of molecular data, and not with an objective image based method as explored here.

Furthermore, we established that if an objective, multi-spectral classification of known species is desired and only one growth substrate is feasible, then using CYA as a growth medium will give a better separation than using YES.

Acknowledgements

The author would like to thank CMIS, CSIRO, Australia for access to VoiR, and in particular Harri Kiiveri for advice and useful discussions during my stay there as a visiting researcher.

Chapter 12

Conclusion

Several examples of large p (no. of input variables), small n (no. of observations) problems were presented and the statistical challenges for such problems were outlined. It was illustrated that dimension reduction in some form is necessary to obtain solutions which are generalizable. At the same time it is of interest to have models which are sufficiently rich to answer the question hand. This trade-off between over- and underfitting is strongly related to the bias-variance trade off which was also explained in the thesis. In the present thesis the questions at hand have been either to classify data into categories or to predict a continuous output variable. For all data examples labelled observations existed, i.e. the output variable was known and the models could therefore be built using this *a priori* knowledge in a supervised setting.

Three novel regularization methods were presented for supervised classification in high-dimensional spaces: Sparse discriminant analysis (SDA), sparse mixture discriminant analysis (SMDA) and orthogonality constrained support vector machine (OC-SVM).

The first two (SDA and SMDA) give sparse solutions which means that only a few of the input variables contribute to the solution. The sparsity was induced using a regularization term of the ℓ_1-norm of the parameter estimates. This was added jointly with a shrinkage term for greater generalization in form of the ℓ_2-norm of the parameter estimates. The regularizations were added to both linear

discriminant analysis and mixture discriminant analysis giving the flexibility of handling both linear separations and also non-linear separations.

The third (OC-SVM) adds *a priori* information of pairing between observations to the support vector machine. This give solutions with less variation than traditional support vector machine solutions. Additionally, small improvements in classification rates were observed for classification of ear canal impressions into whether they were acquired with open or closed mouth.

Additionally, novel applications of sparse regressions (the elastic net) to the medical, concrete and food industries via mutlispectral images for objective and automated systems were presented. These were estimation of moisture content in sand used for concrete, quality assessment of colors of deep fried frozen vegetables, and estimation of severity scores of psoriasis lesions.

The thesis emphasizes the use of *a priori* information available in data such as assumptions of redundant and irrelevant features, pairing of observations, labelled data and control measurements. The latter two give rise to supervised analyses. When such information is exploited it is of great importance to give accurate measures of the prediction errors of the models. This can be obtained by probably separating validation where model parameters and dimensions are tuned from the test of the model, i.e. a separate test data set is utilized which have not been included in any of the previous modelling steps to report prediction errors.

For future research bayesian ways of tuning parameters, and semi-supervised analyses are of interest. The bayesian way of tuning parameters give estimates of e.g. sparsity which do not depend on data and with the scarce data available for many of the problems presented here this is an advantage. Cross-validation is still a good way of comparing different models and reporting prediction errors once parameters have been chosen. Furthermore, massive databases are becoming publicly available as measure apparatus get better, faster and cheaper all the time. This means that more data will become available. Data which in general is unlabelled and this can be exploited in unsupervised settings.

Appendix A

Multiplicative updates for the LASSO

Authors: Morten Mørup[1] and Line H. Clemmensen[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

Published in proceedings *2007 IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING : MLSP2007*, 2007, p. 33-38.

A.1 Abstract

Multiplicative updates have proven useful for non-negativity constrained optimization. Presently, we demonstrate how multiplicative updates also can be used for unconstrained optimization. This is for instance useful when estimating the least absolute shrinkage and selection operator (LASSO) i.e. least squares minimization with L_1-norm regularization, since the multiplicative updates (MU) can efficiently exploit the structure of the problem traditionally solved using quadratic programming (QP). We derive two algorithms based on MU for the LASSO and compare the performance to Matlabs standard QP solver as well as the basis pursuit denoising algorithm (BP) which can be obtained from
www.sparselab.stanford.edu. The algorithms were tested on three benchmark bio-informatic datasets: A small scale data set where the number of observations is larger than the number of variables estimated ($M < J$) and two large scale microarray data sets ($M \gg J$). For small scale data the two MU algorithms, QP and BP give identical results while the time used is more or less of the same order. However, for large scale problems QP is unstable and slow. both algorithms based on MU on the other hand are stable and faster but not as efficient as the BP algorithm and converge slowly for small regularizations. The benefit of the present MU algorithms is that they are easy to implement, they bridge multiplicative updates to unconstrained optimization and the updates derived monotonically decrease the cost-function thus does not need any objective function evaluation. Finally, both MU are potentially useful for a wide range of other models such as the elastic net or the fused LASSO. The Matlab implementations of the LASSO based on MU can be downloaded from Mørup and Clemmensen (2007).

A.2 Introduction

Multiplicative updates were introduced to solve the non-negative matrix factorization (NMF) problem, i.e. factor analysis with non-negativity constraints imposed on all variables Lee and Seung (1999, 2000). This has recently been extended to semi-NMF, i.e. where the parameters under consideration are non-negative while the data in itself is unconstrained Ding et al. (2006); Sha et al. (2002). We will presently advance the multiplicative updates to unconstrained

optimization, i.e. problems where the parameters can both take positive and negative values. We demonstrate, that these types of updates are useful to solve least squares problems with L_1-norm penalty also referred to as the LASSO Tibshirani (1996).

The least absolute shrinkage and selection operator (LASSO), is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values also named the L_1-norm of the coefficients Tibshirani (1996), i.e.

$$\boldsymbol{\beta} = \arg\min\{\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\|_F^2\} \quad s.t. \quad \sum_m |\boldsymbol{\beta}_m| \leq t, \tag{A.1}$$

which is equivalent to the minimization

$$\boldsymbol{\beta} = \arg\min\{\frac{1}{2}\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\|_F^2 + \lambda \sum_m |\boldsymbol{\beta}_m|\}. \tag{A.2}$$

That is, there is a one to one correspondence between t and λ Tibshirani (1996); Chen et al. (1999). LASSO has connections to soft-thresholding of wavelet coefficients, forward stagewise regression, and boosting methods Efron et al. (2004) and forms a framework to solve the Basis Pursuit Shaobing and Donoho (1994); Guigue et al. (2005) with noise (Basis Pursuit Denoising) Chen et al. (1999). The attractive property of the L_1-norm is that it penalizes the non-sparsity of $\boldsymbol{\beta}$ without violating the convexity of the optimization problem. Furthermore, the L_1-norm is known to mimic the behavior of the L_0 norm, i.e. to attain as many zero elements as possible **?** giving the simplest and often also the most parsimonious solution to account for the data.

The equivalent minimization problems given in equation (A.1) and (A.2) have been solved by quadratic programming (QP). Since $|\boldsymbol{\beta}_m|$ cannot be handled by regular QP the problem has been recast in the non-negative variables $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ such that $\boldsymbol{\beta}_m = \boldsymbol{\beta}_m^+ - \boldsymbol{\beta}_m^-$. Then, the LASSO can be stated in standard QP form by $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ -\mathbf{X} \end{bmatrix}$ and $\widetilde{\boldsymbol{\beta}} = [\boldsymbol{\beta}^+, \boldsymbol{\beta}^-]$ subject to the constraint $\widetilde{\boldsymbol{\beta}} \geq \mathbf{0}$. We will currently explore the structure of this reformulated problem to form two algorithms for the LASSO based on multiplicative updates. Using multiplicative updates has the following benefits:

1. The non-negativity constraint of $\widetilde{\boldsymbol{\beta}}$ can naturally be enforced.
2. The fact that $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ -\mathbf{X} \end{bmatrix}$ can be used to avoid doubling the size of the problem compared to standard QP-solvers.

3. The algorithm based on multiplicative updates is easy to implement, has low computational cost per iteration and is proven to monotonically decrease the cost-function.

4. The multiplicative updates form a general optimization framework which can potentially be used for a wide range of problems.

A.3 Method

Multiplicative updates (MU) were introduced in Lee and Seung (1999, 2000) to solve the non-negative matrix factorization (NMF) which corresponds to

$$\mathbf{Y} \approx \boldsymbol{\beta}\mathbf{X}, \tag{A.3}$$

where $\mathbf{Y} \in \Re_+^{I\times J}$, $\boldsymbol{\beta} \in \Re_+^{I\times M}$ and $\mathbf{X} \in \Re_+^{M\times J}$ are all non-negative. This was extended to semi-NMF Ding et al. (2006) where $\mathbf{Y} \in \Re^{I\times J}$ and $\mathbf{X} \in \Re^{M\times J}$ i.e. for $\boldsymbol{\beta}$ non-negativity constrained while $\mathbf{Y}$ and $\mathbf{X}$ are unconstrained. Given a cost function $C(\boldsymbol{\beta})$ over the non-negative variables $\boldsymbol{\beta}$, define $\frac{\partial C(\boldsymbol{\beta})^+}{\partial \boldsymbol{\beta}_{i,m}}$ and $\frac{\partial C(\boldsymbol{\beta})^-}{\partial \boldsymbol{\beta}_{i,m}}$ as the positive and negative part of the derivative with respect to $\boldsymbol{\beta}_{i,m}$. Then the multiplicative update has the following form

$$\boldsymbol{\beta}_{i,m} \leftarrow \boldsymbol{\beta}_{i,m} \left(\frac{\frac{\partial C(\boldsymbol{\beta})^-}{\partial \boldsymbol{\beta}_{i,m}}}{\frac{\partial C(\boldsymbol{\beta})^+}{\partial \boldsymbol{\beta}_{i,m}}} \right)^{\alpha} . \tag{A.4}$$

A small constant $\varepsilon = 10^{-9}$ is added to the numerator and denominator to avoid division by zero or forcing $\boldsymbol{\beta}$ to zero. If the gradient is positive $\frac{\partial C(\boldsymbol{\beta})^+}{\partial \boldsymbol{\beta}_{i,m}} > \frac{\partial C(\boldsymbol{\beta})^-}{\partial \boldsymbol{\beta}_{i,m}}$, hence, $\boldsymbol{\beta}_{i,m}$ will decrease and vice versa if the gradient is negative. Thus, there is a one-to-one relation between fixed points and the gradient being zero. α is a "step size" parameter that potentially can be tuned. Notice, when $\alpha \to 0$ only very small steps in the negative gradient direction are taken. The attractive property of multiplicative updates is that they automatically enforce non-negativity while given values of α have been proven to monotonically decrease various cost functions. For NMF the Kullback-Leibler divergence and least squares cost functions are monotonically decreased for $\alpha = 1$ Lee and Seung (2000) while semi-NMF based on least squares as defined in Ding et al. (2006) is monotonically decreased for $\alpha = 0.5$ Ding et al. (2006). Another form of multiplicative updates for semi-NMF is given in Sha et al. (2002) derived in the framework of quadratic programming with non-negativity constraints.

Presently, we will demonstrate that multiplicative updates can also be used for unconstrained optimization, that is $\mathbf{Y} \in \Re^{I\times J}$, $\boldsymbol{\beta} \in \Re^{I\times M}$ and $\mathbf{X} \in \Re^{M\times J}$ are

unconstrained. Notice, this problem can be trivially solved by matrix inverses. However, it is relevant to solve the problem by multiplicative updates when constraints such as sparseness by the L_1-norm is imposed since a closed form solution no longer exists. Furthermore, such constraints are traditionally imposed when the problem is over complete ($M \gg J$) and matrix inverses become unstable. Without loss of generality we will consider $\boldsymbol{\beta} \in \Re^{1\times M}$. We now have the LASSO problem as stated in equation (A.2)

$$\boldsymbol{\beta} = \arg\min\{\frac{1}{2}\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\|_F^2 + \lambda \sum_m |\boldsymbol{\beta}_m|\}. \tag{A.5}$$

If $\boldsymbol{\beta}$ is unconstrained the gradient of the L_1-term, i.e. $P = \lambda \sum_m |\boldsymbol{\beta}_m|$, gives $\frac{\partial P}{\partial \boldsymbol{\beta}} = \lambda \cdot sign(\boldsymbol{\beta})$ ($\boldsymbol{\beta} \neq 0$) such that the contribution from the constraint gives a step of same length regardless of the value of $\boldsymbol{\beta}$. Consequently, for large scale sparse problems oscillations around zero of small elements of $\boldsymbol{\beta}$ makes a simple gradient search get stuck in small step-sizes in order to keep decreasing the cost function. However, by reformulating the problem in the variables $\boldsymbol{\beta}_m = \boldsymbol{\beta}_m^+ - \boldsymbol{\beta}_m^-$ and constraining $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ to be non-negative elements can no longer cross zero. Furthermore, the non-differentiability at $\boldsymbol{\beta} = 0$ is no longer a concern as $\boldsymbol{\beta}$ only goes to zero from one direction. Presently, non-negativity can naturally be enforced by multiplicative updates. Consider again the reformulated LASSO problem cast in the non-negative variables $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ to be solvable by QP

$$C_{LASSO} = \frac{1}{2}\|\mathbf{Y} - \widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{X}}\|_F^2 + \lambda \sum_m \widetilde{\boldsymbol{\beta}}_m, \tag{A.6}$$

where $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ -\mathbf{X} \end{bmatrix}$ and $\widetilde{\boldsymbol{\beta}} = [\boldsymbol{\beta}^+, \boldsymbol{\beta}^-]$. The gradient of the cost function is given by

$$\frac{\partial C_{LASSO}}{\partial \widetilde{\boldsymbol{\beta}}} = -(\mathbf{Y} - \widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{X}})\widetilde{\mathbf{X}}^T + \lambda \mathbf{1} \tag{A.7}$$

Notice further, that

$$\mathbf{Y} - \widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{X}} = \mathbf{Y} - (\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)\mathbf{X} = \mathbf{Y} - \boldsymbol{\beta}\mathbf{X} \tag{A.8}$$

$$\mathbf{Y}\widetilde{\mathbf{X}}^T = [\mathbf{Y}\mathbf{X}^T, -\mathbf{Y}\mathbf{X}^T] \tag{A.9}$$

$$\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T = \begin{bmatrix} \mathbf{X}\mathbf{X}^T & -\mathbf{X}\mathbf{X}^T \\ -\mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{X}^T \end{bmatrix}. \tag{A.10}$$

Using multiplicative updates (MU) as given in equation (A.4), we now get (for $\boldsymbol{\beta} \in \Re^{I\times M}$)

$$\boldsymbol{\beta}_{i,m}^+ \leftarrow \boldsymbol{\beta}_{i,m}^+ \sqrt{\frac{([\mathbf{Y}\mathbf{X}^T]^+ + \boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^- + \boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^+)_{i,m}}{([\mathbf{Y}\mathbf{X}^T]^- + \boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^+ + \boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^-)_{i,m} + \lambda}}$$

$$\boldsymbol{\beta}_{i,m}^- \leftarrow \boldsymbol{\beta}_{i,m}^- \sqrt{\frac{([\mathbf{Y}\mathbf{X}^T]^- + \boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^+ + \boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^-)_{i,m}}{([\mathbf{Y}\mathbf{X}^T]^+ + \boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^- + \boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^+)_{i,m} + \lambda}}$$

where $[\mathbf{M}]^+$ and $[\mathbf{M}]^-$ denotes the positive and negative part of $\mathbf{M}$. Based on the approach of Sha et al. (2002) the following multiplicative updates (MUqp) can also be derived

$$\boldsymbol{\beta}^+_{i,m} \leftarrow \boldsymbol{\beta}^+_{i,m} \frac{-\mathbf{P}_{i,m} + \sqrt{\mathbf{P}^2_{i,m} - 4(\boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^+)_{i,m}(\boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^-)_{i,m}}}{2(\boldsymbol{\beta}^+[\mathbf{X}\mathbf{X}^T]^+)_{i,m}}$$

$$\boldsymbol{\beta}^-_{i,m} \leftarrow \boldsymbol{\beta}^-_{i,m} \frac{-\mathbf{R}_{i,m} + \sqrt{\mathbf{R}^2_{i,m} - 4(\boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^+)_{i,m}(\boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^-)_{i,m}}}{2(\boldsymbol{\beta}^-[\mathbf{X}\mathbf{X}^T]^+)_{i,m}}$$

where $\mathbf{P} = -\mathbf{Y}\mathbf{X}^T - \boldsymbol{\beta}^-\mathbf{X}\mathbf{X}^T + \lambda\mathbf{1}$ and $\mathbf{R} = \mathbf{Y}\mathbf{X}^T - \boldsymbol{\beta}^+\mathbf{X}\mathbf{X}^T + \lambda\mathbf{1}$. A proof, that the first type of updates (MU) monotonically decrease the cost function is given in the Appendix, see section 5. An equivalent proof for the second type of updates (MUqp) follows directly from Sha et al. (2002). Thus, the algorithms formed by the updates above do not need to evaluate the objective function. Notice, for both algorithms all that is needed in memory is the precalculated values $\mathbf{Y}\mathbf{X}^T$ and $\mathbf{X}\mathbf{X}^T$ while each iteration requires computations of size $\boldsymbol{\beta}\mathbf{X}\mathbf{X}^T$. Consequently, the computational complexity is given as $\mathcal{O}(IM^2)$. Furthermore, the problem is in theory convex and therefore not prone to local minima. However, one problem is to estimate when the algorithm has converged. Presently, we defined the convergence as a small relative change in $\boldsymbol{\beta}$ less than 10^{-8} or when 20000 iterations had been reached. To speed up the algorithm, we used active sets to disregard very small elements in $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$. Furthermore, for $\lambda = 0$ the activity of $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ is arbitrary for fixed difference, i.e $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$. Thus, if an element in $\boldsymbol{\beta}$ changed infinitesimally between each iteration the complete activity of this element was placed in either $\boldsymbol{\beta}^+$ or $\boldsymbol{\beta}^-$ depending on the sign of the element in $\boldsymbol{\beta}$ to further reduce the problem size.

A.4 Results and Discussion

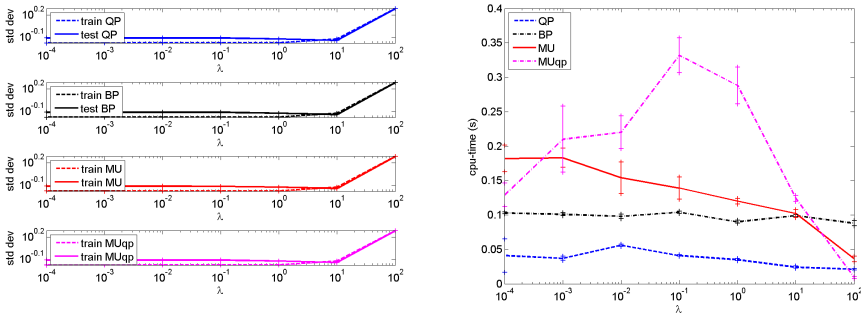
We tested the two types of multiplicative updates presently derived for the LASSO against the standard solver in Matlab for quadratic programming (QP) and the basis pursuit denoising (BP) algorithm described in Chen et al. (1999) which is available from www.sparselab.stanford.edu. Three data sets were considered: One small scale and two large scale problems.

A.4.1 Small scale data set ($M < J$)

The first example is a well known study performed by Stamey et al. (1989) also used as an example in Hastie et al. (2009), where $M < J$. The study examined the correlation between the level of specific prostate antigen and 8

clinical measures ($M = 8$). The clinical measures were taken on 97 men ($J = 97$) who were about to receive a radical prostatectomy.

For the data set, we see that the solutions of MU, MUqp, QP and BP are equivalent in standard error (given as $\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\mathbf{Y}_j - (\widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{X}})_j)^2}$), see figure A.1 (a). The cpu-time usage is of same order for MU, MUqp, BP and QP although QP is slightly faster than the other three, see figure A.1 (b).



(a) Std dev (b) Cpu-time

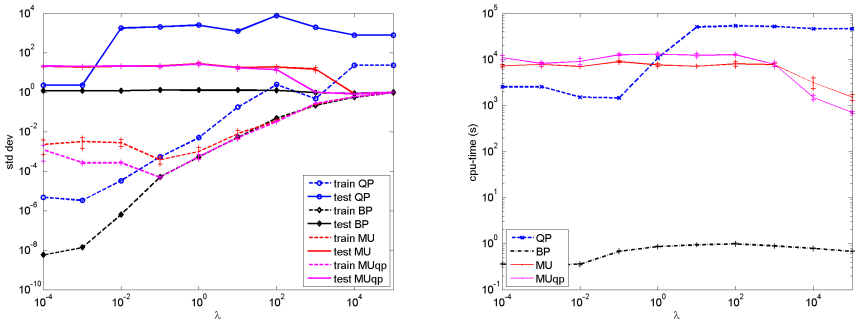
Figure A.1: The standard deviation and the cpu-time as a function of λ for the prostate cancer data. The solutions found by QP, MU, MUqp and BP are identical while the time-usage is of more or less the same order. The time usage of MU and MUqp reduces for large values of λ due to occurence of zero elements which can be disregarded in the updates. The error bars denotes the standard deviation of the mean of 10 runs.

A.4.2 Large scale data sets ($M \gg J$)

The two large scale data sets consist of microarray data taken from Pochet et al. (2004) of studies performed by Alon et al. (1999) and Hedenfalk et al. (2001) of colon data and breast cancer data, respectively. The microarrays contain expressions of 2000 and 3226 genes.

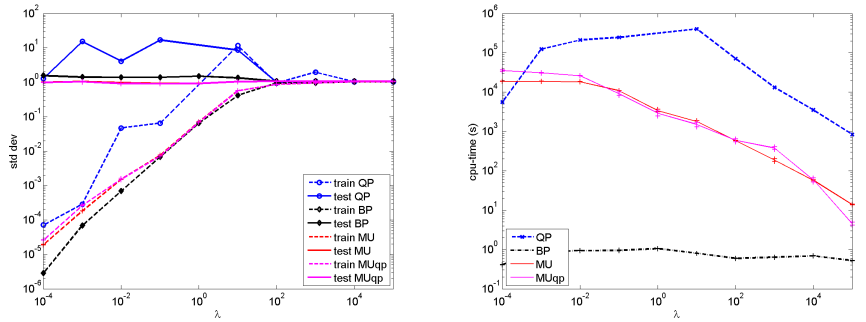
The first data set represents a study of the gene expression for 40 tumor and 22 normal colon tissues. The samples were divided into a training set with 13 normal samples and 27 tumor samples and a test set with 9 normal samples and 13 tumor samples.

The second data set considers gene expressions for carriers of BRCA1 mutation,



(a) Std dev (b) Cpu-time

Figure A.2: The standard deviation and the cpu-time as a function of λ for the colon cancer data. While QP is unstable and slow, MU and MUqp are more stable than QP. However, for small values of λ the multiplicative updates are slower than QP and does not fully converge. For large values of λ MU and MUqp is faster than QP and the solutions of MU and MUqp are equivalent to those obtained by BP. The error bars denotes the standard deviation of the mean of 3 runs, due to the large computational time for QP only one run of QP has been included.



(a) Std dev (b) Cpu-time

Figure A.3: The standard deviation and the cpu-time as a function of λ for the breast cancer data. The same tendencies are observed as for the colon cancer data in figure A.2. Namely, For small values of λ MU and MUqp have not fully converged. Furthermore, QP is again unstable and slow. MU and MUqp is faster and for large values of λ equivalent to BP in quality of solutions obtained. The error bars denotes the standard deviation of the mean of 3 runs, due to the large computational time for QP only one run of QP has been included.

BRCA2 mutation, and sporadic cases of breast cancer. Here, we will consider the separation of BRCA1 mutations from the tissues with BRCA2 mutations or sporadic mutations. The training set consists of 4 samples with BRCA1 mutations and 10 without. The test set consists of 3 samples with BRCA1 mutations and 5 without.

The results obtained from the colon data set as well as the breast cancer data set are given in figure A.2 and figure A.3, respectively. For small values of λ both MU and MUqp have not fully converged however for large values of λ the solutions are equivalent to BP. Finally, QP is unstable and have problems finding the minima regardless of the values of λ.

A.5 Conclusion and future work

The present algorithm for the LASSO based on two types of multiplicative updates performed equally well for small scale problems as QP and BP. However, for large scale over complete problems BP was much faster than both QP, MU and MUqp. For large values of λ BP, MU and MUqp had same quality of solutions but for low values MU and MUqp did not converge. While QP was unstable for large scale problems this was neither the case for MU, MUqp nor BP. Although, multiplicative updates suffer from slow convergence when λ is small they are simple and easy to implement and clearly outperform QP for large scale problems. However, they are not as good as state of the art algorithms such as the BP algorithm Chen et al. (1999). The present multiplicative updates were based on two different approaches, Ding et al. (2006); Sha et al. (2002). Despite their different nature their performances were for the present analysis very similar.

Other algorithms for the LASSO than the present QP and BP exist, see for instance Osborne et al. Osborne et al. (2000). Also, algorithms not based on directly minimizing the LASSO cost for a specific value of λ such as least angle regression selection (LARS) Efron et al. (2004) and the Homotopy method Drori and Donoho (2006); Osborne et al. (2000) have recently been proposed. However, these algorithms are based on successively introducing or removing variables rather than directly minimizing the cost-function for a specific value of λ hence do not directly compare to the present algortihms for the LASSO based on multiplicative updates.

The multiplicative updates based on equation (A.4) is a general framework to solve non-negativity constrained problems and can easily be generalized to other cost-functions and additional constraints. Presently, we demonstrated

that multiplicative updates can be used for unconstrained minimization where $\boldsymbol{\beta}$ takes both positive and negative values and how this could be used to form two simple algorithm for the LASSO. Recently, the LASSO has been advanced to the so called "elastic net" which apart from a L_1-norm penalty has an additional L_2-norm penalty on $\boldsymbol{\beta}$. This encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together and improves the stability in the $M \gg J$ case for small values of λ Zou and Hastie (2005). Furthermore, the LASSO has been advanced to the fused LASSO where the L_1-norm is imposed on both the coefficients and their successive differences. This encourages local constancy of the coefficient profile and also improves stability in the $M \gg J$ case Tibshirani and Saunders (2005). It should be possible to advance the present multiplicative updates to both accommodate the elastic net as well as the fused LASSO. This will be the focus of future work. Furthermore, in Zhang et al. (2006) it was demonstrated that multiplicative updates easily can accommodate missing values - this might be relevant to consider when modeling data using the LASSO. Hence, it is our strong belief that the present multiplicative methods can be extended to form simple algorithms for a wide range of data as well as models.

A.6 APPENDIX: Proof of convergence for MU $\alpha = 0.5$

The proof is based on the use of an auxiliary function Lee and Seung (2000) and follows closely the proofs for the convergence of semi-NMF given in Ding et al. (2006). Briefly stated, an auxiliary function G to the function F is defined by: $G(\mathbf{B}, \mathbf{B}') \geq F(\mathbf{B})$ and $G(\mathbf{B}, \mathbf{B}) = F(\mathbf{B})$. If G is an auxiliary function then F is non-increasing under the update $\mathbf{B} = \arg\min_{\mathbf{B}} G(\mathbf{B}, \mathbf{B}')$.

Let $\mathbf{B} \in \Re_+^{I \times M}$. In Ding et al. (2006) the following relations are proven to hold

$$
\begin{aligned}
Tr(\mathbf{B}[\mathbf{X}\mathbf{X}^{\mathbf{T}}]^{+}\mathbf{B}) &\leq \sum_{i,m} \frac{([\mathbf{X}\mathbf{X}^T]^{+}\mathbf{B}')_{i,m}\mathbf{B}_{i,m}^2}{\mathbf{B}'_{i,m}} \\
Tr(\mathbf{B}[\mathbf{X}\mathbf{X}^{\mathbf{T}}]^{-}\mathbf{B}) &\geq \sum_{i,m,m'} [\mathbf{X}\mathbf{X}^T]^{-}_{m,m'}\mathbf{B}'_{i,m}\mathbf{B}_{i,m'} \\
&\quad (1 + log\frac{\mathbf{B}_{i,m}\mathbf{B}_{i,m'}}{\mathbf{B}'_{i,m}\mathbf{B}'_{i,m'}}) \\
Tr(\mathbf{B}[\mathbf{Y}]^{-}) &\leq \sum_{i,m}[\mathbf{Y}]^{-}_{i,m}\left(\frac{\mathbf{B}_{i,m}^2 + \mathbf{B}'^2_{i,m}}{2\mathbf{B}'_{i,m}}\right)
\end{aligned}
$$

$$
\begin{aligned}
Tr(\mathbf{B}[\mathbf{Y}]^{+}) &\geq \sum_{i,m}[\mathbf{Y}]^{+}_{i,m}\mathbf{B}_{i,m}(1+\log\frac{\mathbf{B}_{i,m}}{\mathbf{B}'_{i,m}}) \\
\mathbf{B}_{i,m} &\leq \frac{\mathbf{B}^2_{i,m}+\mathbf{B}'^2_{i,m}}{2\mathbf{B}'_{i,m}}
\end{aligned}
$$

The present LASSO costfunction is given as:

$$
C(\widetilde{\boldsymbol{\beta}}) = \frac{1}{2}\|\mathbf{Y}-(\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{-})\mathbf{X}\|^2_F + \lambda\sum_{i,m}(\boldsymbol{\beta}^{+}_{i,m}+\boldsymbol{\beta}^{-}_{i,m})
$$

$$
\begin{aligned}
&= \frac{1}{2}Tr(\mathbf{Y}\mathbf{Y}^{\mathbf{T}}) \\
&+ \frac{1}{2}Tr((\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{-})([\mathbf{X}\mathbf{X}^{T}]^{+}-[\mathbf{X}\mathbf{X}^{T}]^{-})(\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{-})^{T}) \\
&- 2Tr((\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{-})([\mathbf{X}\mathbf{Y}^{T}]^{+}-[\mathbf{X}\mathbf{Y}^{T}]^{-})) \\
&+ \lambda\sum_{i,m}(\boldsymbol{\beta}^{+}_{i,m}+\boldsymbol{\beta}^{-}_{i,m})
\end{aligned}
$$

Using the upper bounds on positive contributions and lower bounds on negative contributions given before, an auxiliary function for $G(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}')$ is derived. Minimizing this function with respect to $\widetilde{\boldsymbol{\beta}}$ we obtain the multiplicative updates with $\alpha = 0.5$.

Appendix B

A comparison of dimension reduction methods with applications to multi-spectral images of sand used in concrete

---

Authors: Line H. Clemmensen[1] and Michael E. Hansen[1] and Bjarne K. Ersbøll[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

Online first version available in *Machine Vision* and *Applications*, 2009.

B.1 Abstract

This paper presents a comparison of dimension reduction methods based on a novel machine vision application for estimating moisture content in sand used to make concrete. For the application in question it is very important to know the moisture content of the sand so as to ensure good-quality concrete. In order to achieve a continuous in-line approach for the concrete mixing, digital image analysis is used. Multi-spectral images, consisting of nine spectral bands in the visible and near infrared (NIR) range, were acquired. Each image consists of approximately 9 million pixels. Five different sand types were examined with 20-60 images for each type. To reduce the amount of data, features were extracted from the multi-spectral images; the features were summary statistics on single bands and pairs of bands as well as morphological summaries. The number of features (2016) is high in relation to the number of observations and, therefore, dimension reductive methods are needed. Furthermore, speed, which is an important consideration, is aided by the use of a small number of variables. On top of that, fewer dimensions tend to give more robust results. Two traditional statistical methods for dimension reduction (forward selection and principal components) combined with ordinary least squares and one sophisticated chemometrics algorithm (genetic algorithm - partial least squares) are compared to the recently proposed least angle regression - elastic net (LARS-EN) model selection method.

B.2 Introduction

Concrete plays an increasingly important role in modern society. It is a composite material which is made up of a filler and a binder. The binder "glues" the filler together to form a synthetic conglomerate. The constituents used for the binder are cement and water, while the filler can be a fine or coarse aggregate. For quality concrete a proper water-cement ratio is essential. Especially, when the aggregate is sand or finer particles, an unspecified amount of water is always present. When adding extra water to the concrete, this latent water must be accounted for in order to obtain the optimal ratio. Current methods build on electrical sensors which need to be in contact with the sample in order to measure the moisture content (Razek, 1989; Carver, 1952), or on neutron-sondes which are designed to be used near the surface of the sample (Walker,

1982). Here, digital imaging is considered to obtain an in-line registration of the moisture content of the sand.

Mixture models are traditionally used for analyzing spectral data such as geological data, which to our knowledge come closest to the sand data presented here, see e.g. Berman et al. (1999). However, in general such spectral data sets are hyper-spectral rather than multi-spectral as here. The low spectral resolution and a high image resolution (1035×1380 pixels) makes it inadequate to apply mixture models to the present data. It becomes difficult to tell background and noise from actual features in the low resolution spectral curves, and the many pixels result in a high computational load. Furthermore, we are not interested in direct pixel estimates of the moisture content but rather in an overall estimate for the entire image/sample. Therefore, it was decided to sum up the spectral information in each image using summarizing statistics before modelling the moisture content. The multi-spectral images give rise to a large number of features (summarizing statistics), p.

Traditional multivariate statistical methods are adequate in situations with few variables relative to the number of observations ($n \gg p$). Unfortunately, the same methods are not applicable in most cases where the situation is reversed as is the case here, i.e. there are more variables than observations. In order to achieve feasible computation times and robust results, it is usual to perform some sort of dimension reduction of the variables (König, 2000). A traditional approach has been dimension reduction by e.g. principal components followed by regression analysis (Lee et al., 2001), or perhaps feature selection as in forward selection (Chan and Lee, 2002; Fodor and Kamath, 2001). Although simple and relatively robust, methods like these obviously suffer from sub-optimalities. In the first case there is no guarantee that the selected principal components represent the relevant information, in the second case n variables (where n is the number of observations) at most can be included in the model before it saturates. Methods like all subsets are computationally impractical due to the large number of features.

In the chemometrics literature, but also in other fields such as genetics, partial least squares (PLS) is often used for dimension reduction as an alternative to principal component analysis because it directs the choice of latent variables in accordance with the response variable. Recently research in these fields has focused on the use of feature selection in combination with PLS to improve the prediction accuracy and further reduce the complexity of the models (Leardi, 2000). Genetic algorithms (GA) are a popular class of feature selection techniques used in conjunction with PLS (Bu et al., 2007; Leardi, 2000). These are a class of optimization procedures which model biological reproduction procedures (Goldberg, 1989). They consist of three steps which are recycled: Selection, crossover, and mutation. The algorithm considered here is based on

100 independent, very short GA runs, in order to reduce the risk of overfitting (Leardi and Gonzalez, 1998).

Newer methods have recently been suggested that integrate the data compression and variable selection in one step. The most recent, known as LARS-EN (Least Angle Regression - Elastic Net), was suggested by Zou and Hastie (2005). LARS-EN regularizes the OLS (Ordinary Least Squares) solution by adding constraints to the 1-norm (Lasso) and the 2-norm (Ridge regression) of the coefficients. This method can perform both regularization and variable selection, or one of these depending on the choise of parameters. This makes it very useful, particularly when the number of variables is much larger than the number of observations.

Finally, cross-validation has proven advantageous as an evaluation technique with regard to parameter tuning (Hastie et al., 2009; Duda et al., 2001), and will thus also be used here.

The rest of the paper is organized as follows: Section 2 presents the sand data, section 3 summarizes the algorithms and methodology used in the paper, section 4 states the results and compares the dimension reduction techniques, and finally section 5 sums up the results in a discussion and states possible future research.

B.3 Data

Five types of sand with different geographical origins were examined in this experiment. Buckets of 10 L with sand and water were mixed in order to attain one of eight required nominal moisture levels. Three samples of small amounts of sand were then taken from each bucket and placed in petri dishes. A digital image of each petri dish was acquired by a multi-spectral camera. After imaging, the moisture content in each sample was measured by placing each sample in an oven that dried out the sample. The amount of vaporized water was measured as the difference between the weight of wet and dry sand, as in Blees (1989). Sampling was conducted so that:

- For sand types 1, 3, and 5 there were three different distributions of grain size: Fine, medium, and coarse, based on the most common grain size in the sample.
- For sand types 2 and 4 there was only one distribution of grain size: Medium.

- The experiments were conducted with up to eight different levels of moisture content. The required nominal moisture levels were: 0%, 1.25%, 2.5%, 3.75%, 5%, 6.25%, 7.5%, and 8.75%-moisture. The standard deviation of the difference between the samples and the required nominal levels was 0.8% moisture.
- Zero, three, six, nine, or twelve repetitions were performed for each set of parameters. Within each group of samples taken from the same bucket, the average standard deviation was 0.1% moisture.

The samples with a required moisture level of 0% were dried at over 100°C. This is not a realistic situation and it causes an abrupt change in the appearance of the samples, which were therefore disregarded in the further analysis.

For sand types 2 and 4 three samples were taken for every moisture level, yielding a total of 21 samples for each sand type. For sand types 1 and 5 three samples were taken for every second moisture level (i.e. for 2.5%, 5%, and 7.5%) at the fine and coarse grain distribution, and three samples for every moisture level at the medium grain distribution except for the 5%-moisture level where 6 samples were taken. In total there were 42 samples for sand types 1 and 5, respectively. For sand type 3 three samples were likewise taken for every second moisture level at the fine and coarse grain distribution. Additionally, at the medium grain distribution 3, 9, 3, 12, 3, 9, and 3 samples were taken for the 7 moisture levels, respectively, yielding a total of 59 samples to analyze for sand type 3.

B.3.1 Digitization

The images of the sand samples were digitized at the Department of Informatics and Mathematical Modelling (IMM), Technical University of Denmark (DTU) with a very accurate 9 channel digital camera system called Videometer Lab[1]. The Videometer Lab is illustrated in Figure B.3 and consists of an integrating sphere illumination (an Ulbricht sphere) combined with a carefully calibrated digital camera. The sphere has a diameter of 36 cm. The inside of the sphere is covered with a matte titanium paint to create optimum light conditions. Light is brought into the system through light emitting diodes (LEDs) at the rim of the sphere, giving the sample a diffuse illumination. The Videometer camera system generates images with a radiometric resolution of 8 bits/pixel for each color-channel. The result is stored as 9 floating-point images, one for each spectral band, with a spatial resolution of 1035×1380 pixels (more than 1 million pixels in each spectral band). The nine spectral bands acquired are listed in Table

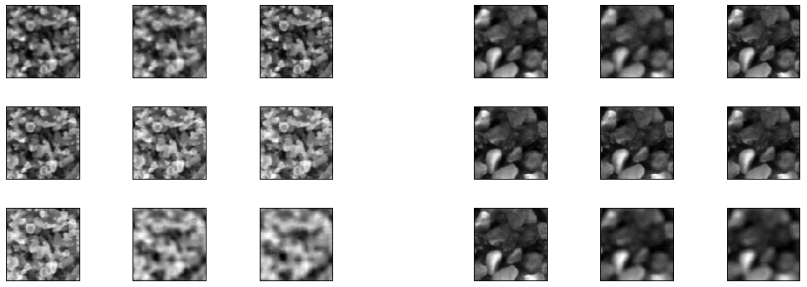
[1] http://www.videometer.com

B.1. For two of the samples the corresponding 9 spectral images are illustrated in Figure B.1. The mean intensity of the spectral images as a function of the wavelength is illustrated in Figure B.2. Note the distinct difference in spectral profile between the two samples.

Color	Ultra blue	Blue	Cyan	Green	Amber	Orange	Red	NIR	NIR
428	472	503	515	592	612	630	875	940

Wavelength (nm)

Table B.1: The wavelengths of the nine spectral bands and a description of the colors corresponding to these wavelengths.



(a) Sample 80 (b) Sample 138

Figure B.1: The nine spectral images for two 100×100 pixels sections of the sand samples. Sample 80: Type 2, medium distribution of grain curve, 2.5% moisture. Sample 138: Type 3, coarse distribution of grain curve, 7.5% moisture.

In the following, each image is considered an observation unless otherwise noted. Later, we will analyze whether this is reasonable by considering the variance of the moisture estimates for smaller areas in the images.

B.3.2 Features

The features that were extracted from each multi-spectral image were summarizing statistics aiming at reducing the vast amount of information present. The

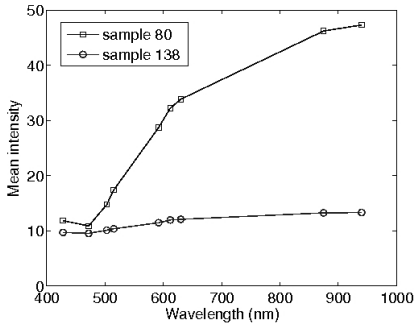
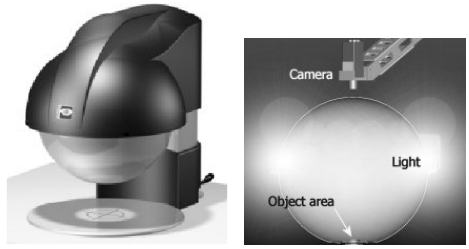


Figure B.2: Illustration of the mean intensity as a function of the spectra for the two samples: 80 and 138.



(a) Camera system (b) Illumination

Figure B.3: The camera system and a cross section of the sphere illustrating the illumination.

1st, 5th, 10th, 30th, 50th, 70th, 90th, 95th, and 99th percentiles (9 summary features) were evaluated of:

- The nine original spectral images ($9 \times 9 = 81$)
- The logarithm of these original spectra ($9 \times 9 = 81$)
- The pairwise differences between all the observations in the original spectra ($9 \times 36 = 324$)
- The pairwise products of all the observations in the original spectra ($9 \times 36 = 324$)
- The pairwise ratios between all the observations in the nine original spectra ($9 \times 36 = 324$)
- The morphological opening and the morphological closing of the nine spectral bands ($9 \times 2 \times 9 = 162$). The structuring element was a disk with radius 5.

The number of features for each subset of images is stated in parentheses. Furthermore, scale-spaces were considered in order to incorporate knowledge of the various grain sizes present for each sand type; see Lindeberg (1994) for more information on scale-spaces. Standardized images, meaning images with zero mean and standard deviation one, were first constructed from the mean images over the nine spectral bands. The scale-spaces were constructed by filtering each standardized image with a rotationally symmetric Gaussian low-pass filter with standard deviations (σ) 0, 1, 2, 5, 10, 15, 20, 25, and 30, and a rectangular structuring element of size $1 \times \sigma$ (9 scale-spaces, and 36 pairwise differences between scale-spaces). The standard deviation, mean, kurtosis, and skewness of the scale-spaces ($4 \times 9 = 36$) and of the pairwise differences between the scale-spaces ($4 \times 36 = 144$) were calculated. Nearest neighbor interpolation were used to construct images of size 1, 0.9, 0.8, 0.6, 0.4, and 0.2 (6 size fractions) times the size of the original scale-spaces and of the pairwise differences between scale-spaces. Additional features for the scale-spaces were: The mean and standard deviation of the gradient of the size fractions of the scale-space images ($2 \times 6 \times 9 = 108$) and of the pairwise differences between scale-spaces ($2 \times 6 \times 36 = 432$). The size fractions were constructed by nearest neighbor interpolation. This amounted to a total of 2016 features.

The logarithms of the measured moisture levels were used as the response variables, as the variation is higher for the high moisture levels than for the low moisture levels.

B.4 Methods

The traditional approach to dimension reduction is to combine either forward selection or principal components with regression analysis. A more sophisticated approach performs feature selection using a genetic algorithm and then partial least squares on the subset with a further selection of the number of components entered in the model. The recently developed LARS-EN (Least angle regression - elastic net) method performs variable selection with data compression and regression analysis in one algorithm.

B.4.1 Forward selection

Forward selection (FS) is a popular variable selection technique for including variables in e.g. a *general linear model* (GLM) (Hastie et al., 2009; Rencher, 2002). The technique starts by evaluating a model containing only a constant. Then the variable with the largest partial correlation with the response variable is chosen. The F-value of the test that the coefficient of this variable is significantly different from zero at an α-level is computed. If it is significantly different from zero the variable is included in the model and the next variable is chosen. If, on the other hand, it can be assumed to be zero, the procedure is stopped. Typically $\alpha = 5\%$ is used. Alternatively, cross-validation can be used to evaluate the number of variables which should be included in the regression model.

B.4.2 Principal component analysis

Principal component analysis (PCA) is a data decomposition technique where the principal components (PCs) of a multivariate data set are mutually orthogonal linear combinations of the original variables. The PCs are composed to explain the largest possible amount of variance in the data, hence, the first PC has the largest variance, the second PC the second largest variance, and so on (Jolliffe, 1986). Since the variance depends on the scale of the original variables, the data is commonly standardized to have zero mean and variance one. The directions of the principal components are the eigenvectors of the data dispersion matrix. In many cases, the first few PCs explain most of the variance of data and are therefore sufficient for further analyses. However, the principal components do not always provide information relevant for the purpose of the analyses.

B.4.3 Genetic algorithm - partial least squares

Genetic algorithm - partial least squares (GA-PLS) consists of feature selection via genetic algorithms and data decomposition via partial least squares. PLS uses the dependent variable to direct the orthogonal variables which are computed, unlike PCA which simply chooses the directions according to the variance in the independent variables (Bastien et al., 2005). Combined with feature selection which further reduces the complexity of the models as well as the predictive ability (Leardi, 2000), GA-PLS is a more sophisticated model selection technique, see e.g. Bu et al. (2007). Genetic algorithms simulate the biological mechanisms of reproduction (Goldberg, 1989). Since they are based on random effects it is customary to run them several times to check their robustness. Here we have used the Matlab PLS-genetic algorithm toolbox provided by Riccardo Leardi. It uses 100 independent, short GA runs to reduce the risk of overfitting; for further details on the algorithm see Leardi and Gonzalez (1998); Leardi (2000). The PLS-genetic algorithm toolbox makes use of 5-fold cross validation, and only returns test error rates. However, a test using a *fitness* score of the average percentage of explained variance when the dependent variable is randomly permuted is provided to check for overfitting.

B.4.4 Least angle regression - elastic net

The *least angle regression - elastic net* (LARS-EN), proposed by Zou and Hastie (2005), regularizes the OLS solution using two constraints, the 1-norm and the 2-norm of the coefficients. These constraints are the ones used in the *least absolute shrinkage and selection operator* (Lasso) (Tibshirani, 1996) and Ridge regression (Hoerl and Kennard, 1970), respectively. The naïve elastic net estimator is defined as

$$\hat{\vec{\beta}} = \text{argmin}_{\beta}\{\|\vec{y} - \mathbf{X}\vec{\beta}\|_2^2 + \lambda_1\|\vec{\beta}\|_1 + \lambda_2\|\vec{\beta}\|_2^2\} \quad , \tag{B.1}$$

where $\vec{y}$ is the dependent variable (in our case the logarithm of the percentage moisture level for each image), $\mathbf{X}$ is a matrix with the independent variables (the summarizing statistics) in the columns, $\vec{\beta}$ denotes the regression parameters, λ_1 is the weight on the Lasso penalty ($\lambda_1 \geq 0$), and λ_2 is the weight on the Ridge penalty ($\lambda_2 \geq 0$). Additionally, $\|\vec{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$, $|\cdot|$ denoting the absolute value, and $\|\vec{\beta}\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. Choosing $\lambda_1 = 0$ yields Ridge solutions, while choosing $\lambda_2 = 0$ yields Lasso solutions.

Let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, then, according to Zou and Hastie (2005), solving (F.5) is

equivalent to solving the optimization problem

$$\hat{\vec{\beta}} = \text{argmin}_\beta\{\|\vec{y} - \mathbf{X}\vec{\beta}\|_2^2\} \quad s.t. \quad (1-\alpha)\|\vec{\beta}\|_1 + \alpha\|\vec{\beta}\|_2^2 \leq t \quad \text{for some } t \quad . \tag{B.2}$$

Without loss of generality we can use the equality to give a relationship between the parameters because variables enter the regression equation sequentially as t increases and at discrete values of t and λ_1; we will explain this in greater detail in the next section.

The function $(1-\alpha)\|\vec{\beta}\|_1 + \alpha\|\vec{\beta}\|_2^2$ is called the elastic net penalty and it is illustrated together with the Ridge and Lasso penalties in Figure B.4. The OLS

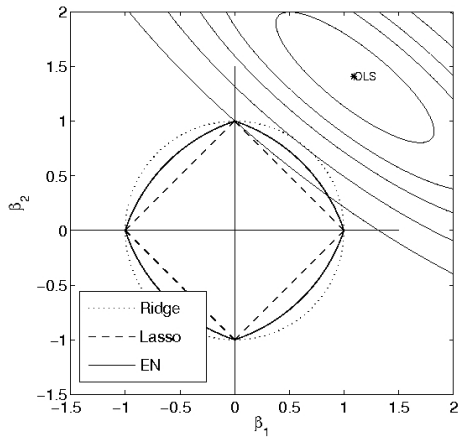


Figure B.4: Example with two parameters: $\vec{\beta} = [\beta_1\ \beta_2]^T$, illustrating the estimation with the constraints: Ridge ($\|\vec{\beta}\|_2^2 \leq 1$), Lasso ($\|\vec{\beta}\|_1 \leq 1$), and elastic net ($(1-\alpha)\|\vec{\beta}\|_1 + \alpha\|\vec{\beta}\|_2^2 \leq 1$ with $\alpha = 0.5$). The OLS solution is found at the center of the contours, and the penalized solutions (Ridge, Lasso, and elastic net) are found where the contours first touch the respective constraints.

solution to a linear problem is also marked in the figure, and the contours of the quadratic function[2] $(\vec{\beta} - \hat{\vec{\beta}})^T\mathbf{X}^T\mathbf{X}(\vec{\beta} - \hat{\vec{\beta}})$ are sketched. The Ridge, Lasso, and elastic net solutions are obtained where the contours first touch the respective constraint. For the Lasso method this is likely to occur at or near a corner (as illustrated in the figure) where one of the coefficients is zero, while for the Ridge regression it is unlikely that one of the coefficients will be zero. The elastic net constraint is somewhere in between Ridge and Lasso.

[2]This function equals the RSS criterion plus a constant, and the contours are centered at the OLS solution (Tibshirani, 1996).

We can transform the naïve elastic net problem into an equivalent Lasso problem on the augmented data (Zou and Hastie, 2005, Lemma 1)

$$\mathbf{X}^* = (1+\lambda_2)^{-1/2} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix} \quad , \quad \vec{y}^* = \begin{bmatrix} \vec{y} \\ \vec{0}_p \end{bmatrix} \quad , \tag{B.3}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, and $\vec{0}_p$ is a length p vector of zeroes. The normal equations, yielding the OLS solution, to this augmented problem are

$$\frac{1}{\sqrt{1+\lambda_2}} \left(\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I}_p^T \mathbf{I}_p\right) \hat{\vec{\beta}}^* \quad = \quad \mathbf{X}^T \vec{y} \quad . \tag{B.4}$$

We see that $\frac{1}{\sqrt{1+\lambda_2}} \hat{\vec{\beta}}^*$ is the scaled Ridge regression estimate with parameter λ_2. Hence, performing Lasso on this augmented problem yields an elastic net solution. The *least angle regression* (LARS) model selection method proposed by Efron et al. (2004) can be used with advantage to compute the Lasso solution on the augmented problem. The LARS algorithm obtains the Lasso solution with a computational speed comparable to that obtained using the OLS solution of the full set of covariates.

Lasso chooses at most n variables, where n is the number of observations, before the solution deteriorates and the method sets all coefficients to nonzero. Since we are interested in variable selection this might be limiting and, therefore, LARS-EN is considered. Furthermore, groups of variables can enter at the same time with the LARS-EN algorithm, unlike e.g. the Lasso, Ridge regression, and forward selection methods.

The algorithm uses the LARS implementation with the Lasso modification as described in the following section. Hence, we have the parameter λ_2 to adjust, but also the number of iterations for the LARS algorithm can be used. The larger λ_2, the more weight is put on the Ridge constraint. The Lasso constraint is weighted by the number of iterations. A small number of iterations corresponds to a high value of λ_1 (resulting in a small number of non-zero loadings), and vice versa.

B.4.5 Least angle regression

The least angle regression (LARS) model selection method (Efron et al., 2004) finds the predictor most correlated with the response, takes a step in this direction until the correlation is equal to another predictor, then it takes the equiangular direction between the predictors of equal correlation (*the least angle direction*) and so forth.

By ensuring that the sign of any non-zero coordinate β_j has the same sign as the current correlation between the residual $\vec{r}$ and the independent variable $\vec{x}_j$ [3], the LARS method yields all Lasso solutions. This result is obtained by differentiating the Lagrange version of the Lasso problem. For further details see Efron et al. (2004).

Summing up, for each iteration in the LARS algorithm one (or more) variable(s) enters the regression equation exactly when the correlation between this variable and the current residual is equal to the correlation between the current residual and the variables which are currently in the regression equation. In the same way, variables only enter the regression equation, i.e. have a non-zero loading, at discrete values of λ_1, and it is only these *jumps* that we are interested in as the Lasso constraint is used as a variable selection technique. Additionally, variables in LARS can be dropped again due to the Lasso sign constraints on β.

B.4.6 Cross-validation

We use *leave-one-out* cross-validation (CV) error estimates to decide the number of variables to include in the model (for the cases: FS, PCA, and GA-PLS), or to choose the set of regularizing parameters to use in the model (for the LARS-EN case) (Hastie et al., 2009; Duda et al., 2001; Rencher, 2002). A full set of n models each based on $n-1$ observations (images) is built by leaving out the i'th observation (image) from the i'th model. The estimate of the j'th observation in the i'th model is denoted $\hat{y}_{(i)j}$. For $j = i$ this is an estimate of the left-out observation (the test-set). For $j \neq i$ this is the usual plug-in estimate of the training set observations. The error estimates of the n 1-observation test sets $(y_i - \hat{y}_{(i)j}, j = i)$ and the n sets of $n-1$ plug-in errors for the training sets $(y_i - \hat{y}_{(i)j}, j \neq i)$ are used to compare the four methods (Duda et al., 2001). We use the standard deviation of the leave-one-out errors to measure the test error, $e_{test} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i)i})^2}$. We use the standard deviation of the training errors to measure the training error, $e_{train} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} v_{(i)}}$, where $v_{(i)}$ is the estimated error variance in the i^{th} model, $v_{(i)} = \frac{1}{n-1}\sum_{j\neq i}(y_j - \hat{y}_{(i)j})^2$. The advantages of leave-one-out CV are many. One is that the error rate is approximately unbiased (Hastie et al., 2009). A second is that the leave-one-out error rates are comparable. The test errors of the models can be compared by calculating the jack-knife estimate of the variance of the test errors, and a statistical t-test of the hypothesis of equal test errors for the two methods can be performed (Duda et al., 2001). The third advantage is that the number

[3] In LARS $\vec{y}$ is assumed centered, and $\mathbf{X}$ is assumed centered and normalized so each variable has unit length which means that the correlations simply are given by $\vec{r}^T\mathbf{X}$.

of observations is rather small in relation to the fairly high sample variation. Hence, in order to avoid large variations of the error estimate, leave-one-out CV seems a good choice (Rencher, 2002). Finally, e_{train} tends to underestimate the prediction error. It also tends to be smaller than e_{test} which is why the latter is usually considered a more honest estimate of the prediction error.

B.4.7 Comparison of the Four Methods

Although the model selection will be performed off-line and only the prediction is to be performed in-line, computational speed is always desirable. LARS-EN is computationally faster than GA-PLS, FS and PCA. The entire elastic net regularization path is computed with the effort of a single OLS fit (Zou and Hastie, 2005), but in practice not all of the path is computed as the number of iterations is used for regularization. A disadvantage is that two regularization parameters are to be adjusted in LARS-EN. However, if the ranges of the parameters are known, the computations are limited. A disadvantage of GA-PLS is the random effect, which means that several GA runs should be performed in order to ensure that robust models are obtained. However, the method is still considerably faster than FS.

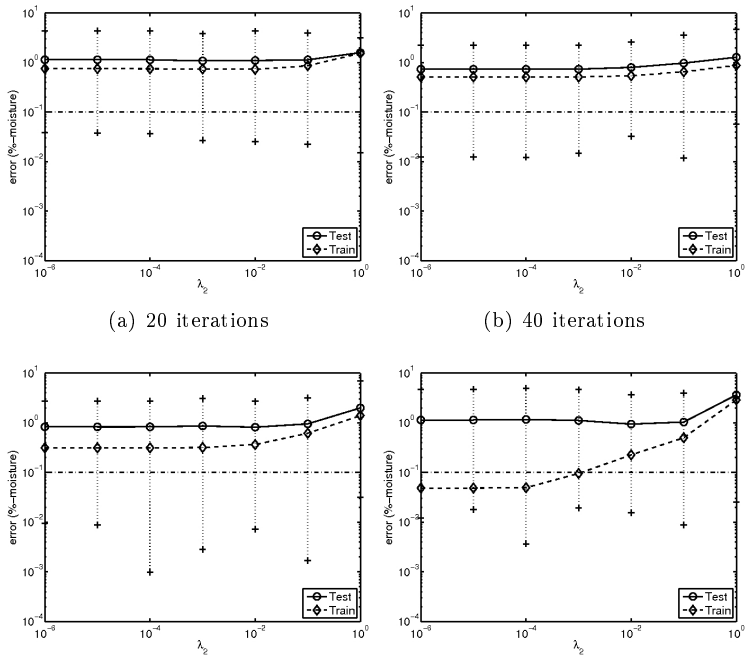
One of the most important assets of a model selection technique is its ability to provide low error rates from a subset of features. The fact that the selected model only relies on a subset of features can make the in-line prediction real-time. An additional advantage is that models based on subsets of features in general are more robust with regards to error rates as they do not tend to overfit and therefore give better prediction accuracy.

While FS and LARS-EN are directly comparable, they are not directly comparable with PCA and GA-PLS. PCA involves linear combinations of all the independent variables, and then a selection approach of the principal components. GA-PLS first involves a feature selection, and then, on the reduced feature set, a data decomposition approach (similar to PCA) with further selection of the number of components.

B.5 Results

The results obtained with the four methods are listed in Table B.2. An example of the cross-validation and performance of LARS-EN is illustrated in Figure B.5. The best choice is (b) where the lowest mean error for the test set is observed,

and the average number of active variables included in the leave-one-out models is 10. After 40 iterations the lowest test error is observed, and in addition to this the training and test errors are closest.



(a) 20 iterations (b) 40 iterations

(c) 80 iterations (d) 160 iterations

Figure B.5: Illustration of the mean test error as a function of λ_2 in the leave-one-out CV for sand type 1 using different numbers of iterations. The minimum mean test errors in the four cases are: 1.1%-, 0.7%-, 0.8%- and 0.9%-moisture, respectively, corresponding to $\lambda_2 = 10^{-3}$, $\lambda_2 = 10^{-4}$, $\lambda_2 = 10^{-2}$, and $\lambda_2 = 10^{-2}$. $e = 10^{-1}$ is marked with a dash-dotted line. The minimum and the maximum test errors observed in the cross-validation are also marked in the graphs. They provide the range of the test errors.

For forward selection, c.f. Table B.2, the test error (e_{test}) is considerably larger than the training error (e_{train}) in more than half of the instances, i.e. the training data is overfitted. The overfitting is observed even though the number of variables included is smaller than the number of observations. For some of the sand types this method yields low errors of the prediction error, despite the overfitting. Additionally, the statistical tests comparing the test errors show that the e_{test} of FS is significantly different, and larger, at a 5% level of significance

than those of both PCA and LARS-EN for sand types 2 and 3. One advantage of FS is that few variables are included in the models which in particular is an advantage in an in-line production due to the smaller computational effort.

Principal component analysis does not tend to overfit data as forward selection does, c.f. Table B.2. However, the test errors for PCA are higher than for forward selection in two out of five cases, but this difference is only significant, and larger, at a 5% level than that of FS and LARS-EN for sand type 5. The disadvantage of PCA is that it includes all variables in each component in the solution. This is particularly a disadvantage in an in-line production because all spectral bands must be acquired, and all summary statistics must be computed before calculating the principle component scores for the regression.

The models selected with LARS-EN do not tend to overfit data like FS and PCA, c.f. Table B.2, i.e. the test errors of the training and the test data are close. In addition to that, the test errors are never statistically larger at a 5%-level of significance than those for FS and PCA. LARS-EN often includes more variables in the selected models (sometimes many more) than FS, and also more variables than PCA includes components. However, due to the coefficient shrinkage ensured by the Ridge weight, λ_2, this does not lead to overfitting. Furthermore, the standard deviations of the test errors for LARS-EN are often smaller (and never larger) than for FS and PCA. On top of that, compared to PCA only a subset of features are included in the models, an advantage in an in-line production. Finally, LARS-EN is comparable to GA-PLS in the test error rates in 4 out of 5 cases, and additionally the models include fewer variables than those of GA-PLS.

Looking at the fitness scores for GA-PLS, we see that is should be safe to use GA-PLS on the data sets of sand types 1, 3, and 5 ($fitness < 10$, cf. Leardi (2000)), whereas the relatively few samples for sand types 2 and 4 means that a slight overfitting is likely. The GA-PLS procedure was repeated five times and the models were fairly robust with regard to the error estimates.

For LARS-EN, the features most often selected have been features from pairwise relations between two of the original spectral images. Which of the original spectral images are of interest varies from sand type to sand type. Furthermore, one or more of the scale-space features are selected for sand types 1, 3, and 5 which contain more than one grain distribution, but not for sand type 2 and 4 which only contain one grain distribution. In addition to this, features which include information from the two NIR bands are frequently included in the models.

Finally, we will comment on the standard deviation of the estimated moisture contents when we consider only sub regions of the images. Table B.3 lists the

Method	Type	e_{train}	e_{test}	$\sigma[e_{test}]$	λ_2	ite	N_v	N_c	$fitness$
FS	1	0.3	0.6	0.6	-	-	7	-	-
PCA	1	0.8	0.7	0.6	-	-	2016	5	-
GA-PLS	1	-	0.4	-	-	132	86	8	7
LARS-EN	1	0.5	0.7	0.5	10^{-4}	41	10	-	-
FS	2	0.1	0.6	0.6	-	-	7	-	
PCA	2	0.3	0.3	0.3	-	-	2016	3	
GA-PLS	2	-	0.3	-	-	52	158	2	12
LARS-EN	2	0.3	0.4	0.3	10^{-1}	129	109	-	
FS	3	0.4	0.8	0.7	-	-	14	-	
PCA	3	0.3	0.5	0.5	-	-	2016	27	
GA-PLS	3	-	0.7	-	-	162	136	7	5
LARS-EN	3	0.4	0.6	0.5	10^{-3}	83	30	-	
FS	4	10^{-4}	0.4	0.35	-	-	16	-	
PCA	4	0.2	0.3	0.2	-	-	2016	9	
GA-PLS	4	-	0.3	-	-	31	58	2	12
LARS-EN	4	0.1	0.3	0.2	10^{-3}	47	21	-	
FS	5	0.3	0.3	0.3	-	-	3	-	
PCA	5	0.4	0.5	0.4	-	-	2016	12	
GA-PLS	5	-	0.4	-	-	40	19	6	8
LARS-EN	5	0.3	0.4	0.3	10^{-3}	18	9	-	

Table B.2: The table lists the average error of the training (e_{train}) and the test (e_{test}). In addition, the jack-knife estimate of the standard deviation of the test error is given ($\sigma[e_{test}]$). The estimates are listed for the four methods: Forward selection (FS), PCA, GA-PLS, and LARS-EN for each of the five sand types. λ_2 and ite are the regularization parameters chosen; in GA-PLS ite is the number of evaluations. N_v is the number of selected variables, and N_c is the number of selected components in the leave-one-out models. $fitness$ is the average percent of variance explained in cross validation with random permutations of the dependent variable; this estimates the degree of overfitting when using GA-PLS.

standard deviation of the three models, where each observation respectively corresponds to: One image, one sub image of half size, and one sub image of a quarter size. The standard deviation increases as smaller areas are considered as observations. This emphasizes the fact that the moisture is not homogeneously distributed in the sand samples. In this case the only way to decrease the sampling variance would be to take larger and/or more samples and change the sampling process according to the guidelines of the theory of sampling in Petersen et al. (2005). Preferably, the samples should be taken as the images would be taken in the construction line: One image covers the width of the construction line, and images are taken at suitable time intervals.

Sand type	σ_1	σ_2	σ_4
1	0.5	0.7	0.7
2	0.3	0.3	0.4
3	0.4	0.9	1.1
4	0.1	0.4	0.8
5	0.3	1.4	1.6
Average	0.3	0.8	0.9

Table B.3: σ_m is the standard deviation of the LARS-EN model where each image is split into m sub images. The models have the same parameters λ_2 and *ite* as in Table B.2.

B.6 Conclusion and future work

Four dimension reduction techniques (Forward Selection, Principal Component Analysis, Genetic Algorithm - Partial Least Squares, and Least Angle Regression - Elastic Net) were compared. The comparison was based on results from five data sets of images of different sand types used for concrete where the aim was to estimate the moisture content.

LARS-EN, GA-PLS, and PCA are more robust than FS in terms of not overfitting training data. In LARS-EN, this robustness is provided by the Ridge coefficient shrinkage whereas for GA-PLS and PCA it is provided by the data decomposition. In addition to this, LARS-EN is robust in terms of the test error never being significantly larger than those of FS and PCA at a 5% level of significance. Additionally, the test errors of LARS-EN and GA-PLS are comparable.

Summing up, GA-PLS and LARS-EN are to be preferred for their robustness

with regards to not overfitting and in general providing the lowest test errors. LARS-EN is to be preferred for its low computational effort, and its selection of a small subset of variables.

The sample variations are fairly high, and a further study of their influence is desirable. Only the surface of the sand sample is captured by the images, while the reference measures are conducted for the entire sample. This gives rise to additional sampling variations which might be reduced by imaging several consecutive images of larger samples, close to what would be the case in the construction line. Once such studies are made, an implementation in the construction line is the next aim.

B.7 Acknowledgements

The authors would like to thank the SCC-consortium, in particular the four institutions: Danish Technological Institute, 4K-Beton A/S, Videometer A/S, and the Department of Informatics and Mathematical Modelling at the Technical University of Denmark, for making the analyses of the sand data possible. Also thanks to two referees for their valuable comments.

Appendix C

Sparse Discrminant Analysis

Authors: Line H. Clemmensen[1] and Trevor Hastie[2] and Bjarne Ersbøll[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.
2. Statistics Department, Stanford university, California, U.S.A.

Submitted to *Technometrics*.

C.1 Abstract

Classification in high-dimensional feature spaces where interpretation and dimension reduction are of great importance is common in biological and medical applications. For these applications standard methods such as microarrays, 1D NMR, and spectroscopy have become everyday tools for measuring thousands of features in samples of interest. Furthermore, the samples are often costly and therefore many such problems have few observations in relation to the number of features. Traditionally such data are analyzed by first performing a feature selection before classification. We propose a method which performs linear discriminant analysis with a sparseness criterion imposed such that the classification, feature selection and dimension reduction are merged into one analysis. The sparse discriminant analysis is faster than traditional feature selection methods based on computationally heavy criteria such as Wilk's lambda, and the results are better with regards to classification rates and sparseness. The method is extended to mixtures of Gaussians which is useful when clusters are present within each class, e.g. biological subgroups. Finally, the methods proposed provide low-dimensional views of the discriminative directions.

C.2 Introduction

Linear discriminant analysis (LDA) is a favored tool for supervised classification in many applications due to its simplicity and robustness. Comparison studies show that a large percentage (typically more than 90%) of the achievable improvement in predictive accuracy, over the simple baseline model, is achieved by LDA (Hand, 2006). Furthermore, the computations leading towards LDA provides low-dimensional projections of data onto the most discriminative directions. However, it fails in some situations:

- When the number of predictor variables is high in relation to the number of observations ($p \gg n$).
- When a single prototype per class is insufficient.
- When linear boundaries are insufficient in separating the classes.

The mentioned situations where LDA fails were previously addressed in penalized discriminant analysis (Hastie et al., 1995a) and discriminant analysis by Gaussian mixtures (Hastie and Tibshrani, 1996), see also flexible discriminant and mixture models (Hastie et al., 1995b). However, in some cases where $p \gg n$ these methods are not adequate since both sparseness and feature selection are desired. A low number of nonzero parameters ensures a better interpretation of the model and additionally tends to overfit training data less than nonsparse methods as illustrated with the elastic net and sparse principal components (Zou and Hastie, 2005; Zou et al., 2006).

It is often desirable to perform feature selection in biological or medical applications such as microarrays. In these applications it is essential to identify important features for the problem at hand for interpretation issues and to improve speed by using models with few nonzero loadings as well as fast algorithms.

During the past decade problems in which the number of features is much larger than the number of observations have received much attention (Donoho, 2000; Hastie et al., 2009; Duda et al., 2001). Here we consider classification problems and propose a method for performing robust discriminant analysis. Previously this issue has been addressed by ignoring correlations between features and assuming independence in the multivariate Gaussian model (naive Bayes) (Bickel and Levina, 2004). We will focus on imposing sparseness in the model (Donoho, 2000) in line with models such as lasso and the elastic net (Tibshirani, 1996; Efron et al., 2004; Zou and Hastie, 2005). The sparseness is imposed by an ℓ_1-norm on the parameters like in Efron et al. (2004); Tibshirani (1996); Trendafilov and Jolliffe (2007). However, also imposing an ℓ_2-norm like in the elastic net (Zou and Hastie, 2005) ensures that good solutions are also obtained when the number of selected features exceeds the number of variables.

The introduction of a sparseness criterion is well known in the regression framework (Tibshirani, 1996; Efron et al., 2004; Zou and Hastie, 2005; Zou et al., 2006) and we shall therefore consider computing the projections for LDA by optimal scoring which computes the projections (discriminant directions) by regression (Hastie et al., 1995a; Ye, 2007; Grosenick et al., 2008; Leng, 2008). Furthermore, the optimal scoring framework allows for an extension to mixtures of Gaussians (Hastie and Tibshrani, 1996).

The paper is organized as follows. Section two describes the sparse LDA and sparse mixture discriminant analysis algorithms, introducing a modification of the elastic net algorithm to include various penalizing matrices. Section three briefly describes shrunken centroids regularized discriminant analysis and sparse partial least squares which are used for comparison. Section four illustrates experimental results on a small illustrative shape based data set of female and male silhouettes and on four high-dimensional data sets: A microarray data set,

spectral and chemical identification of fungi plus classification of fish species based on shape and texture features. We round off with a discussion in section five.

C.3 Methodology

Linear discriminant analysis (LDA) is a classification method which assumes that the variables in each of the k classes are normally distributed with means μ_j, $j = 1, ..., k$ and equal dispersion Σ (see e.g. Hastie et al. (2009)). Reduced-rank LDA has the ability to provide low-dimensional views of data of up to at most $k-1$ dimensions (Hastie et al., 2009). These views, also called discriminant directions, are furthermore sorted such that the direction discriminating the classes most is first and so forth. The at most $k-1$ directions, β_js are p-dimensional vectors, where p is the number of predictor variables. They are chosen to maximize the variance between classes and minimize the variance within classes subject to being orthogonal to each other. Hence, we maximize the between groups sums of squares, $\Sigma_B = \sum_{j=1}^{k} (\mu_j - \mu)(\mu_j - \mu)^T$ (where μ is the mean of all groups) relative to the within sums of squares, $\Sigma_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \mu_j)(X_{ij} - \mu_j)^T$ (Fisher's criterion)

$$\arg\max_{\beta_j} \beta_j^T \Sigma_B \beta_j \tag{C.1}$$

under the orthogonality constraint

$$\beta_j^T \Sigma_W \beta_l = \begin{cases} 0 & l = 1, ..., j-1 \\ 1 & l = j \end{cases} , \tag{C.2}$$

to find the discriminating directions β_j, $j = 1, ..., k-1$.

The methodology section is written following the notation of Penalized Discriminant Analysis (PDA) in Hastie et al. (1995a). PDA replaces the within sums of squares matrix in (C.2) with the penalized term $\Sigma_W + \lambda_2 \Omega$. In order to obtain sparseness in the solution we introduce an extra term which controls the ℓ_1-norm of the parameters β. The ℓ_1-norm has previously proved to be an effective regularization term for obtaining sparseness; see methods such as lasso, elastic net and sparse principal component analysis (Tibshirani, 1996; Zou and Hastie, 2005; Zou et al., 2006). The sparse discriminant criterion then becomes

$$\arg\max_{\beta_j} \beta_j^T \Sigma_B \beta_j - \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \tag{C.3}$$

under the constraint (C.2) with the penalized within sums of squares matrix $\Sigma_{W_p} = \Sigma_W + \lambda_2 \Omega$ replacing Σ_W.

The elastic net proposed by Zou and Hastie (2005) solves a regression problem regularized by the ℓ_2-norm and the ℓ_1-norm in a fast and effective manner. The elastic net is defined as

$$\beta_j^{en} = \arg\min_{\beta_j}(\|y - X\beta_j\|_2^2 + \lambda_2\|\beta_j\|_2^2 + \lambda_1\|\beta_j\|_1) \quad . \tag{C.4}$$

As the sparse discriminant criterion is also regularized by an ℓ_2-norm and an ℓ_1-norm penalty it seems advantageous to rewrite the criterion as a regression type problem in order to use the elastic net algorithm for solving SDA.

LDA was rewritten in Hastie et al. (1995a) as a regression type problem using optimal scoring; see also the relation between optimal scoring, canonical discriminant analysis and linear discriminant analysis in the appendix C.7.1. The idea behind optimal scoring is to turn categorical variables into quantitative variables using a linear operator (The categorical variables will here be encoded as $\{0,1\}$ *dummy* variables). Optimal scoring assigns a score, θ_{ji} for each class i and for each parameter vector β_j. The optimal scoring problem is defined as

$$(\hat{\theta}, \hat{\beta})^{os} = \arg\min_{\theta,\beta} n^{-1}\|Y\theta - X\beta\|_2^2 \tag{C.5}$$

$$s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad , \tag{C.6}$$

where Y is a matrix of dummy variables representing the k classes, θ is the $k \times q$ matrix of scores, and β is a $p \times q$ matrix where the columns are the β_js.

PDA adds a penalty of $\beta_j^T\Omega\beta_j$ to the optimal scoring problem such that the penalized optimal scoring criterion becomes

$$(\hat{\theta}, \hat{\beta})^{pos} = \arg\min_{\theta,\beta}(n^{-1}\|Y\theta - X\beta\|_2^2 + \lambda_2\|\Omega^{\frac{1}{2}}\beta\|_2^2) \quad , \tag{C.7}$$

s.t. (C.6), where Ω is a symmetric and positive definite matrix. In this paper, a sparseness criterion is added to the penalized optimal scoring criterion via the ℓ_1-norm of the regression parameters β. The normal equations can thus no longer be applied and it is not possible to solve the sparse discriminant analysis (SDA) problem in one regression and one eigenvalue decomposition step as is the case for PDA. We propose an iterative algorithm for solving SDA. Extending the method to mixtures of Gaussians is straightforward in line with Hastie and Tibshrani (1996).

Since the elastic net (Zou and Hastie, 2005) is used in the algorithm we will assume that data are normalized, i.e. the features are transformed to have zero mean and length one. The elastic net algorithm uses the correlation between the dependent variable and the predictors to decide which variable to activate in each iteration. However, it is possible to run the algorithm on raw data which is comparable to performing principal component analysis on the covariance matrix rather than the correlation matrix.

C.3.1 Sparse discriminant analysis by optimal scoring

In this section we introduce constraints to the optimal scoring problem in (C.15) in order to obtain sparseness in the PDA. The score vector θ_j assigns a real number θ_{ji} for each class i, $i = 1, ..., k$. The scored training data $Y\theta$ is an $n \times q$ matrix on which we will regress the matrix of predictors $X_{n\times p}$ to obtain the parameters or directions $\beta_{p\times q}$. This leads to q components of sparse discriminative directions, where $q \leq k-1$ since there are $k-1$ non-trivial directions in the optimal scoring problem. We define sparse optimal scoring as

$$(\theta, \beta)^{sos} = \arg\min_{\theta,\beta} n^{-1}(\|Y\theta - X\beta\|_2^2 + \lambda_2\|\Omega^{\frac{1}{2}}\beta\|_2^2 + \lambda_1\|\beta\|_1) \quad \text{(C.8)}$$

$$s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad , \quad \text{(C.9)}$$

where Ω is a penalization matrix, as introduced in PDA (Hastie et al., 1995a). The ℓ_1-norm introduces sparseness as in lasso or elastic net regularization. In appendix C.7.1 the relation between sparse discriminant analysis (C.3) and sparse optimal scoring (C.8) is given.

For fixed θ we obtain:

$$\beta_j^{sos} = \arg\min_{\beta_j} n^{-1}(\|Y\theta_j - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \quad \text{(C.10)}$$

which for $\Omega = I$ is an elastic net problem. We will later rewrite the elastic net for more general penalty matrices. For fixed β the optimal scores are

$$\theta^{os} = \arg\min_{\theta} n^{-1}\|Y\theta - X\beta\|_2^2 \quad \text{(C.11)}$$

$$s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad .$$

Set $D_\pi = n^{-1}Y^TY$ which is a diagonal matrix of the class proportions. Then the constraint (C.9) can be written as $\theta^T D_\pi \theta = I$ and setting $\theta^* = D_\pi^{\frac{1}{2}}\theta$ we can solve the following problem instead.

$$\hat{\theta}^* = \arg\min_{\theta^*} n^{-1}\|YD_\pi^{-\frac{1}{2}}\theta^* - \hat{Y}\|_2^2 \quad \text{(C.12)}$$

$$s.t. \quad \|\theta^*\|_2^2 = 1 \quad , \quad \text{(C.13)}$$

where $\hat{Y} = X\beta$. This is a balanced Procrustes problem when Y and $\hat{Y}$ have the same dimensions (for $q = k$). As $q \leq k-1$ we pad $\hat{Y}$ with zeros, so that $\hat{Y} = [X\beta \;\; 0]$. The problem can then be solved by taking the svd of $D_\pi^{-\frac{1}{2}}Y^T\hat{Y}$, as described in Elden and Park (1999). However, as we only need to estimate U and V of the svd in order to obtain a solution, and $D_\pi^{-\frac{1}{2}}$ is a diagonal matrix, taking the svd of $Y^T\hat{Y} = USV^T$ suffices, and the solution becomes

$$\hat{\theta}^* = UV^T \Leftrightarrow \quad \text{(C.14)}$$

$$\hat{\theta} = D_\pi^{-\frac{1}{2}}UV^T \quad . \quad \text{(C.15)}$$

By analogy with the PDA case, we use heuristics from suitable normal assumptions as guidelines for producing posterior probabilities and a classifier. As a graphical projection of a predictor vector x we use the set of fits $\beta^T x$. A *nearest class mean* rule, where "nearest" is measured using Σ_{W_p}, is applied in the $q \leq k-1$ reduced-dimensional discriminant subspace to obtain class labels.

C.3.2 Modified elastic net

For generalization, we modify the elastic net algorithm to include an arbitrary penalty matrix Ω rather than the identity. The modified naïve elastic net solution becomes

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|y - X\beta_j\|_2^2 + \lambda_2 \beta_j^T \Omega \beta_j + \lambda_1 \|\beta_j\|_1) \quad . \tag{C.16}$$

We can transform the naive elastic net problem into an equivalent lasso problem on the augmented data (Zou and Hastie, 2005, Lemma 1).

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix} \quad , \quad y^* = \begin{bmatrix} y \\ 0_p \end{bmatrix} \quad . \tag{C.17}$$

The normal equations, yielding the OLS solution, to this augmented problem are

$$\begin{aligned} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix} \hat{\beta}^* &= \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix}^T \begin{bmatrix} y \\ 0_p \end{bmatrix} \Leftrightarrow \\ \left(\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{\Omega}\right) \hat{\beta}^* &= \mathbf{X}^T y \quad . \end{aligned} \tag{C.18}$$

We see that β^* is the Ω-penalized regression estimate with weight λ_2. Hence, performing lasso on this augmented problem yields a modified elastic net solution. Since Ω is symmetric and positive definite, $\sqrt{\Omega}$ always exists. For examples of various penalty matrices Ω and their applications we refer to Hastie et al. (1995a).

C.3.3 Sparse Discriminant Algorithm

The SDA algorithm using optimal scores and the modified elastic net is described in Algorithm 1.

Algorithm 1 Sparse Discriminant Analysis:

1. Initialize $\theta = (k \sum_{j=1}^{k} D_{\pi,\{jj\}})^{-1} I_{1:k-1}$.
2. For $j = 1, ..., q$ solve the modified elastic net problem with fixed θ

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|Y\theta_j - X\beta_j\|_2^2 + \lambda_2 \beta_j^T \Omega \beta_j + \lambda_1 \|\beta_j\|_1) \qquad \text{(C.19)}$$

3. For fixed β and $Y^T \hat{Y} = USV^T$ compute the optimal scores from (C.15).
4. Repeat step 2 and 3 until convergence or a maximum number of iterations is reached.
5. Update β for fixed θ using (C.19), the sparse discriminant directions are now ordered according to the singular values and thereby degree of discrimination.

The sparse discriminant analysis algorithm has a computational effort similar to that of sparse principal component analysis (Zou et al., 2006). It likewise performs an elastic net step and an SVD in each iteration. The elastic net step for $p \gg n$ has the highest computational cost which is in the order of $qO(pnm + m^3)$ where m is the number of nonzero coefficients. This can be massive if p and m are large. However, in general few nonzero coordinates are desired in the mentioned applications, and the algorithm therefore becomes very effective. Additionally, the number of iterations needed to obtain a good solution is generally small.

C.3.4 Sparse mixture of Gaussians

Instead of representing each class by a single prototype we now represent each class by a mixture of Gaussians. We divide each class j into R_j subclasses and define the total number of subclasses $R = \sum_{j=1}^{k} R_j$. To limit the number of parameters we consider a Gaussian mixture model where each subclass has its own mean μ_{jr} and common covariance matrix Σ. Since the single prototype problem is formulated as an optimal scoring problem it is straight forward to extend it to mixtures of Gaussians in line with Hastie and Tibshrani (1996). Instead of using an indicator response matrix Y we use a blurred response matrix $Z_{n \times R}$ which consists of the subclass probabilities, z_{jr} for each observation. Let π_{jr} be the mixing probability within the r^{th} subclass within the j^{th} class, and $\sum_{r=1}^{R_j} \pi_{jr} = 1$. Recall the EM steps of using Bayes' theorem to model Gaussian mixtures. The *estimation* steps of the subclass probabilities, z_{jr} and the mixing

probabilities, π_{jr} are

$$z_{ir} = \frac{\pi_{jr}\exp\{-\frac{(X-\mu_{jr})\Sigma^{-1}(X-\mu_{jr})}{2}\}}{\sum_{r=1}^{R_j}\pi_{jr}\exp\{-\frac{(X-\mu_{jr})\Sigma^{-1}(X-\mu_{jr})}{2}\}} \quad \text{(C.20)}$$

$$\pi_{jr} = \sum_{i\in g_i} z_{ir}, \quad \sum_{r=1}^{R_j}\pi_{jr} = 1 \quad \text{(C.21)}$$

with the *maximization* steps

$$\mu_{jr} = \frac{\sum_{i\in g_i} x_i z_{ir}}{\sum_{i\in g_i} z_{ir}} \quad \text{(C.22)}$$

$$\Sigma = n^{-1}\sum_{j=1}^{k}\sum_{i\in g_i}\sum_{r=1}^{R_j} z_{ir}(x_i-\mu_{jr})(x_i-\mu_{jr})^T \quad . \quad \text{(C.23)}$$

We now write the SMDA algorithm by computing $Q \leq R-1$ sparse directions for the subclasses in the mixture of Gaussians model as described in algorithm 2.

The graphical projections are again given by the sets of fits $x^T\beta$, and a decision rule is given by summing the subclass probabilities calculated using (C.20) for each observation.

C.4 Methods for comparison

Apart from PDA mentioned in the previous section, forward selection (FS; Hastie et al. (2009)) combined with LDA, as well as shrunken centroids regularized discriminant analysis (RDA; Guo et al. (2007)) and sparse partial least squares regression (SPLS; Chun and Keles (2009)) are used for comparisons.

C.4.1 Shrunken centroids regularized discriminant analysis

Shrunken centroids regularized discriminant analysis (RDA) is based on the same underlying model as LDA, i.e. normal distributed data with equal dispersion; Guo et al. (2007). The method regularizes the covariance matrix in LDA,

similar to what is the case for PDA. The regularization of the covariance matrix Σ in RDA is

$$\tilde{\Sigma} = \alpha\hat{\Sigma} + (1-\alpha)I_p \tag{C.27}$$

for some α, $0 \leq \alpha \leq 1$. The singular value decomposition trick is used to compute the matrix inversion for the LDA solution thereby reducing the computational cost from the order of $O(p^3)$ to $O(pn^2)$. The regularization and reduced computational cost makes it suitable for $p \gg n$ problems.

Instead of shrinking the centroids directly it was proposed to shrink $\bar{x}^* = \tilde{\Sigma}\bar{x}$ with the shrinkage parameter $\Delta > 0$, where $\bar{x}$ is the centroid, i.e.

$$\bar{x}^{*'} = \text{sgn}(\bar{x}^*)(|\bar{x}^*| - \Delta)_+ \tag{C.28}$$

which possesses a feature elimination property. However, the false-positive rate for the selected features is high and it is therefore considered a conservative feature selection method.

RDA is based on the same underlying model but uses a different algorithmic approach to overcome the $p \gg n$ problem than SDA. An R package called rda is available from CRAN.

C.4.2 Sparse partial least squares

Sparse partial least squares (SPLS) builds on the partial least squares model (PLS) which assumes that X and Y can be rewritten using basic latent decompositions and thereby obtain a dimension reduction; Chun and Keles (2009). PLS is widely used in chemometrics societies for obtaining dimension reductions in $p \gg n$ problems. SPLS promotes the lasso zero property onto a surrogate direction vector c instead of the original latent direction vector α, while keeping α and c close.

$$\min_{\alpha,c} -\kappa\alpha^T M\alpha + (1-\kappa)(c-\alpha)^T M(c-\alpha) + \lambda_1\|c\|_1 + \lambda_2\|c\|_2 \quad \text{s.t.} \alpha^T\alpha = 1 \tag{C.29}$$

where $M = X^TYY^TX$, $0 \leq \kappa \leq 1$, and λ_2 and λ_1 are the weights on the ridge and lasso constraints, respectively. By using dummy variables to represent the groups with 0s and 1s as Y, PLS and SPLS can be used for classification.

The algorithm uses alternating steps of SVD and elastic net keeping c and α fixed, in the respective steps. The algorithmic approach is thus similar to that of SDA but the underlying model differs. An R package called spls is available from CRAN.

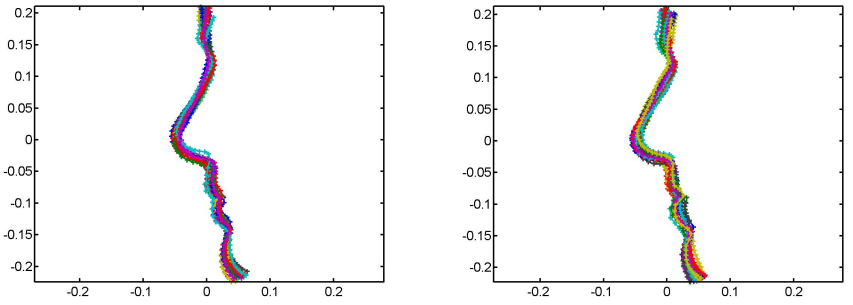
C.5 Experimental results

This section illustrates results on a small data set of shapes from female and male face-silhouettes and on three different high-dimensional data sets: A benchmark high-dimensional microarray data set, a data set based on spectral imaging of *Penicillium* fungi for classification to the species level, a data set with 1D NMRs of three fungal genera for classification to the genus level, and a data set with shape and texture features of three fish species. The number of iterations the algorithms used to obtain good, stable solutions in the following applications were less than 30 in all cases. The parameters for the elastic net as well as for the methods included for comparisons were chosen using leave-one-out cross validation on the training data Hastie et al. (2009). Five- or ten-fold cross validation could also be used, but as we have very few observations in some of the examples we chose to use leave-one-out cross validation. Subsequently, the models with the chosen parameters were tested using the test data. Data was normalized and the penalty matrix $\Omega = I$ unless otherwise mentioned.

C.5.1 Female and male silhouettes

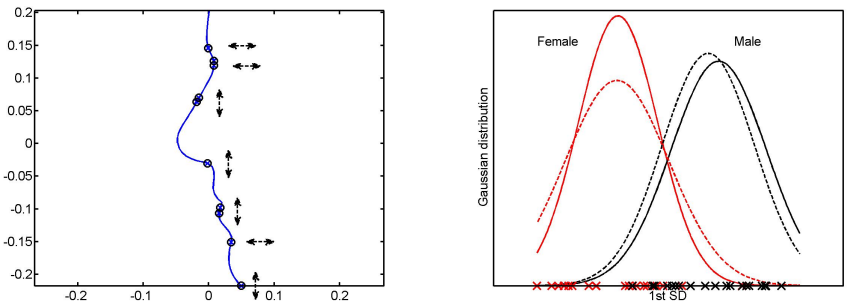
To illustrate the sparse representation of the discriminant directions from SDA we considered a shape based data set consisting of 20 male and 19 female face-silhouettes from adults. A minimum description length (MDL) approach to annotate the silhouettes were used as in Thodberg and Ólafsdóttir (2003), and Procrustes alignment was performed on the resulting 65 MDL landmarks of (x, y)-coordinates. For training, 22 of the silhouettes were used (11 female and 11 male), which left 17 silhouettes for testing (8 female and 9 male). Figure C.1 illustrates the two classes of silhouettes.

Performing leave-one-out cross validation on the training data we selected 10 nonzero features and $\lambda_2 = 10^{-2}$ as parameters for SDA. The SDA results are illustrated in figure C.2. Note, how the few landmarks included in the model were placed near high curvature points in the silhouettes. The training and test classification rates were both 82%. In the original paper (Thodberg and Ólafsdóttir, 2003) a logistic regression was performed on a subset of PCA scores, where the subset was determined by backwards elimination using a classical statistical test for significance. Results were only stated for leave-one-out cross validation on the entire data set which gave a 85% classification rate, see Thodberg and Ólafsdóttir (2003). The SDA model in figure C.2 is easy to interpret compared to a model based on 2-4 principal components each with contributions from all 65 MDL marks. The SDA model points out exactly where the main differences between the two genders are.



(a) Female (b) Male

Figure C.1: The silhouettes and the 65 landmarks for the two groups: Female and male subjects.



(a) Model (b) SD

Figure C.2: Results from SDA on the silhouette data. (a) The mean shape of the silhouettes and the model with the 10 nonzero loadings illustrating which landmarks differ from female to male subjects. The arrows illustrate the directions of the differences. (b) The sparse direction discriminating the classes. The crosses illustrate the observations, the solid curves illustrate the estimated Gaussian distributions of the classes from the training set, and the dashed curves illustrate the estimated Gaussian of the classes from the training and the test set.

C.5.2 Leukemia-subtype microarray

This section considers a high-dimensional benchmark data set from the Kent Ridge Biomedical Data Set Repository (http://sdmc.i2r.a-star.edu.sg/rp/), namely the leukemia-subtype data set published in Yeoh et al. (2002). The study was a genetic analysis of the cancerous cells and aimed at classifying subtypes of pediatric acute lymphoblastic leukemia (ALL). Cancer diseases require fast and correct diagnosis and one way to facilitate this is by microarray analysis. The microarray data set considered here consisted of 12558 genes, 6 subtypes of cancer, 163 training samples and 85 test samples. The six major cytogenetic diagnostic groups in data were: BCR-ABL, E2A-PBX1, Hyperdiploid>50 chromosomes, MLL rearrangement, T-ALL and TEL-AML1. Originally, in Yeoh et al. (2002), data was analyzed in two steps: A feature selection step and a classification step. Furthermore, data were analyzed in a decision tree structure such that one group was separated using an SVM at each tree node. Here, we illustrate the strengths of SDA which performs feature selection, dimension reduction and classification in one step. With only 25 nonzero features, compared to 40 in Yeoh et al. (2002), in each of the 5 discriminant directions classification rates coparable to those in Yeoh et al. (2002) were obtained. The results are summarized in table C.1 and are on non-normalized data for comparison with the original analysis of data. There were 2 misclassified observations in the training set and 3 misclassified observations in the test set. In the latter case all the misclassified observations belonged to the BCR_ABL group but were classified as Hyperdiploid>50.

Figure C.3 illustrates scatter plots of the six groups projected onto the sparse directions obtained by SDA. Note, that each sparse direction separates different groups. This leads to knowledge not only of the separation of all groups, but also of which genes have a different expression level for one subtype of cancer compared to the others, similar to the decision tree structure in the original analysis.

As the confusion mainly is about BCR-ABL there might be more than one genetic mechanism associated with BCR-ABL. Therefore, a mixture model was analyzed where the training observations in BCR-ABL were first split into two subgroups using K-means clustering. The results from SMDA only gave one misclassification, one of the BCR-ABL in the training set was still misclassified as Hyperdiploid>50, but there were no misclassification in the test set.

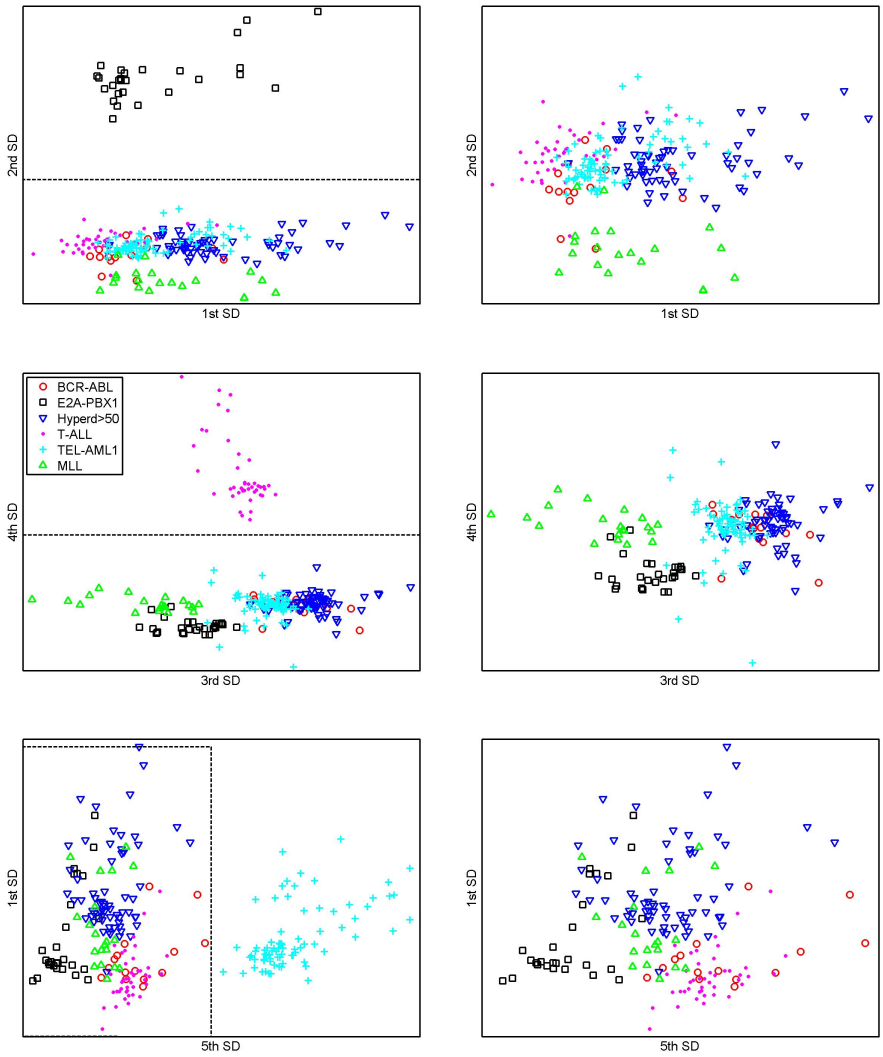


Figure C.3: Sparse discriminant variables in SDA of the Leukemia-subtype data set. The figures to the right are zoom-ins of the areas marked with broken lines in the figures to the left.

Table C.1: Subgroup predictions using SDA with 25 nonzero features in each of the 5 discriminant directions. The ridge weight, $\lambda_2 = 10^{-1}$ as well as the number of nonzero loadings were chosen using leave-one-out cross validation on the training set. The last two columns give the subclass predictions using SMDA where BCR-ABL was modelled using a mixture of Gaussians. The parameters used in SMDA were the same as in SDA.

Group	Train SDA	Test SDA	Train SMDA	Test SMDA
All groups	99%	96%	99%	100%
BCR-ABL	89%	50%	89%	100%
E2A-PBX1	100%	100%	100%	100%
Hyperdiploid>50	98%	100%	100%	100%
T-ALL	100%	100%	100%	100%
TEL-AML1	100%	100%	100%	100%
MLL	100%	100%	100%	100%

C.5.3 Spectral id of fungal species

This section analyzes another high-dimensional data set which considers multi-spectral imaging for objective classification of fungi. Few of the world's fungal species are known today (Hawksworth, 2001) and due to the various useful and toxic mycotoxins they can produce it is of great interest to quickly and accurately classify known species and identify unknown ones. Here, we consider the three *Penicillium* species: *Melanoconodium, polonicum*, and *venetum*. The three species all have green/blue conidia (the spores of the fungi) and are therefore visually difficult to distinguish. It is desirable to have accurate and objective classification of the fungi species as they produce different mycotoxins. Some are very useful to us, such as penicillin while others can be harmful. A visual classification is based on the phenotypes of the species and is in general faster and cheaper than chemical or genetic methods for classification. Using image analysis to perform the classification additionally gives an objective and accurate method which can be reproduced in various laboratories.

For each of the three species, four strains were inoculated on yeast extract sucrose (YES) agar in three replica, in total 36 samples. The data set consisted of 3542 variables extracted from multi-spectral images (Clemmensen et al., 2007) with 18 spectral bands (10 in the visual range, and 8 in the near infrared range). The variables were summary statistics taken from histograms of the fungal colonies in each spectral band, and in each pairwise difference and pairwise multiplication between spectral bands. Table C.2 summarizes the results from reduced-rank PDA, forward selection (FS) based on Wilk's Lambda, and SDA. The data was partitioned into 2/3 which was the training data and 1/3 which was the test data where one of the three repetitions of each strain was

left out for testing. This gave 28 training samples and 12 test samples. In this case the classification rates were not improved, but the complexity of the models was reduced by both SDA and FS. Furthermore, the computational cost of FS based on Wilk's Λ was larger than for SDA. The CPU-time has more than doubled which for just two nonzero loadings is not very excessive but as the number of nonzero loadings increases, the computational cost also increases. Moreover, the two methods: FS and SDA had one of the selected variables in common. Figure C.4 illustrates the sparse discriminant directions in SDA. It is not surprising that the three groups are completely discriminated as they differ in their conidium color which range from green to blue, see Clemmensen et al. (2007). The selected features are thus also percentiles in differences of blue and green spectral bands.

Table C.2: Classification rates from PDA, SDA and forward selection based on Wilk's Λ (FS) combined with LDA on the *Penicillium* data. The Ridge penalty weight was 10^{-6} for PDA and SDA, chosen using leave-one-out cross-validation on the training set. Likewise the number of nonzero loadings was chosen using cross-validation. The covariance matrix in the reduced-rank PDA was ridge regularized since $p >> n$. The chosen parameters for RDA were $\alpha = 0.1$ and $\Delta = 0.2$, and for SPLS there were 5 latent variables with 762 nonzero variables and $\eta = 0.8$. Note, that the computational complexity for forward selection was much larger than for SDA.

Method	Train	Test	Nonzero loadings	CPU-time
PDA	100%	100%	7084	384.3s
FS	100%	100%	2	0.4s
RDA	100%	100%	3502	0.0s
SPLS	100%	100%	3810	0.7s
SDA	100%	100%	2	0.1s

C.5.4 Chemical id of fungal genera

In the previous section we used visual information to classify fungi to the species level. Here we will use chemical information in form of 1D NMR of fungi for classification to the genus level (Rasmussen, 2006). Three genera of fungi were considered: *Aspergillus, Neosartorya*, and *Penicillium*. For each genus there were 5, 2, and 5 species, respectively. There were 71 observations with 4-8 samples of each species. Information from the 950 highest peaks in the NMR data were used as features. Data were logarithmically transformed as differences in peaks with lower intensities seemed to have influence. As the biology gave a hierarchy of subgroups within each genus it seemed reasonable to model each

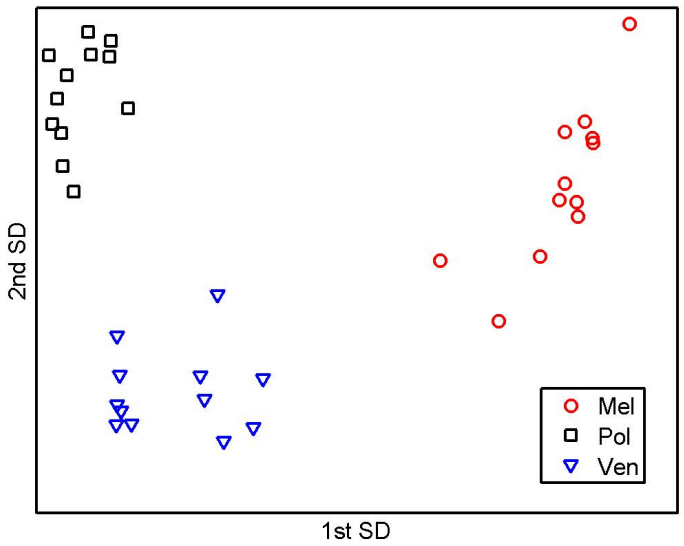


Figure C.4: The *Penicillium* data set projected onto the sparse discriminant directions in SDA.

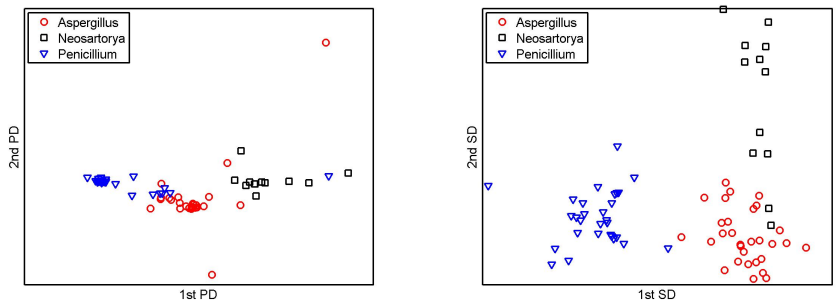
genus as a mixture of Gaussians, i.e. a mixture of species and therefore we tested the SMDA on this data. Table C.3 summarizes the results using PDA, SDA and SMDA on the 1D NMR data. In addition to improved classification rates the sparse methods provided insight in which chemical features that distinguish the fungal genera. Furthermore, the sparse methods gave models with smaller complexity and thereby smaller variance. Consequently, the sparse methods tended to overfit less than the more complex PDA model. Figure C.5 and C.6 illustrate the (sparse) discriminative directions for PDA, SDA, and SMDA. Note, that due to the underlying mixture of Gaussians model, the sparse directions in the SMDA provided knowledge of the separation between genera not only at the genus level but also at the species level.

Table C.3: Errors from PDA, RDA, SPLS, SDA and SMDA on the 1D NMR data. With few nonzero loadings in SDA and SMDA the test classification rates are improved. The Ridge penalty weight is in the range $[10^{-3}, 10^{-1}]$ for the three methods PDA, SDA, and SMDA, and was as well as the number of nonzero loadings chosen using leave-one-out cross validation on the training set. The covariance matrix in the reduced-rank PDA was ridge regularized since $p >> n$. For RDA on normalized data (n) the chosen parameters were $\alpha = 0.99$ and $\Delta = 0.6$, and on the original data (u) they were $\alpha = 0.9$ and $\Delta = 0.5$. For SPLS 8 latent variables including 79 nonzero variables were chosen and $\eta = 0.8$.

Method	Train	Test	Nonzero loadings
PDA	100%	76%	1900
RDA(n)	100%	85%	888
RDA(u)	100%	97%	648
SPLS	100%	88%	632
SDA	97%	91%	10
SMDA	100%	94%	44

C.5.5 Classification of fish species based on shape and texture

Here we consider classification of three fish species: cod, haddock, and whiting. The classification was performed based on shape and texture features. The data taken from Larsen et al. (2009) consisted of 108 fish: 20 cod, 58 haddock, and 30 whiting. The shape of the fish was represented with landmarks based on MDL, as in the example with the male and female silhouettes, see Figure C.7. There were 700 points for the contour of the fish, 300 for the mid line, and 1 for the eye. The shapes were Procrustes aligned to have full correspondence. The texture features were simply the R, G, and B intensity values from color images



(a) PDA (b) SDA

Figure C.5: Discriminant directions in PDA and SDA of the 1D NMR data set. In particular for *Aspergillus* and *Neosartorya* there seems to be subclusters within the genera.

taken with a standard camera under a standardized white light illumination and digitized. They were annotated to the shapes using a Delauney triangulation approach; examples of the texture features are illustrated in Figure C.7. There were a total of 103348 shape and texture features in the data set. In Larsen et al. (2009), a principal component analysis followed by a linear discriminant analysis was performed; resulting in a 76% resubstitution rate. Here, we split data in two: 76 fish for training, and 32 fish for testing. The results are listed in table C.4. In this case, SDA gives the sparsest solution and the best test classification rate. Only one of the whiting was misclassified as haddock.

The sparse discriminant directions are illustrated in Figure C.8. The 1st SD is mainly dominated by blue intensities, this means that cod fish in general are less blue than haddock and whiting around the mid line and mid fin. This is in line with specialists knowledge saying that there is a thin dark line present around the mid line in haddock and whiting whereas this is absent in cod; Larsen et al. (2009). The 2nd SD implies that haddock in general is more blue around the head and tail, less green around the mid line, more red around the tail and less red around the eye, the lower part and the mid line than cod and whiting.

C.6 Discussion

Linear discriminant analysis and classification by mixtures of Gaussians are widely used methods for dealing with supervised classification. In this paper we

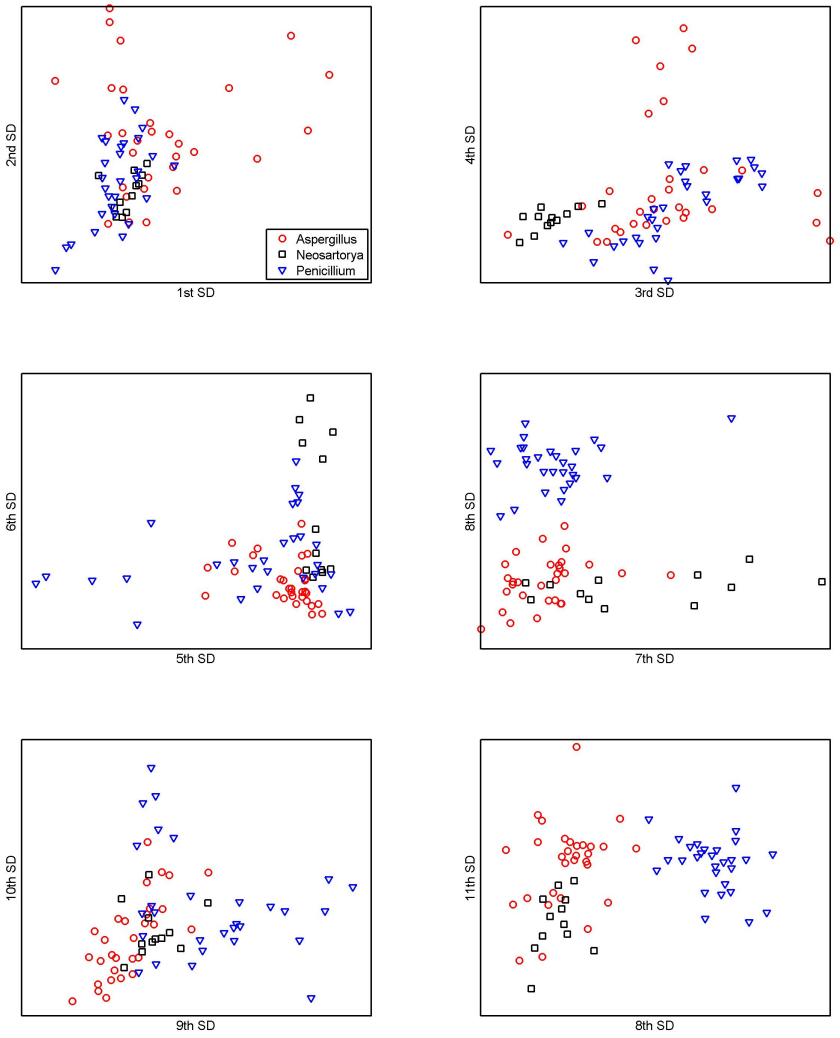


Figure C.6: Sparse discriminant directions in SMDA of the 1D NMR data set. Note how the distribution of each group has changed due to the underlying mixture of Gaussians model. Here, each sparse direction aims at separating one sub group from the remaining.

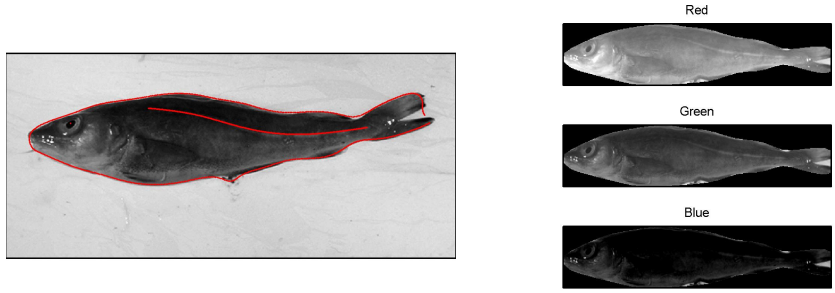


Figure C.7: Illustration of the shape features to the left and the texture features to the right. The shape features are marked as white crosses on a gray scale image. The texture features are the intensities in the red, green and blue bands of the color images.

Table C.4: Errors from SDA, RDA, and SPLS on the shape and texture features from the images of fish. The Ridge penalty weight was 10^{-3} and 30-nonzero loading were included for each of the 2 discriminant directions for SDA, and was as well as the number of nonzero loadings chosen using leave-one-out cross validation on the training set. For RDA on normalized data (n) the chosen parameters were $\alpha = 0.99$ and $\Delta = 0.2$, and on the original data (u) they were $\alpha = 0.99$ and $\Delta = 0.1$. For SPLS 3 latent variables including 105 nonzero variables were chosen and $\eta = 0.9$. The spls package implemented in R could not handle the size of the fish data, and therefore spls was also used preliminarily on subsets of the data to reduce the number of variables for the problem. Note, that RDA performed much better when data was not normalized.

Method	Train	Test	Nonzero loadings
RDA(n)	100%	41%	103084
RDA(u)	100%	94%	103348
SPLS	100%	81%	315
SDA	100%	97%	60

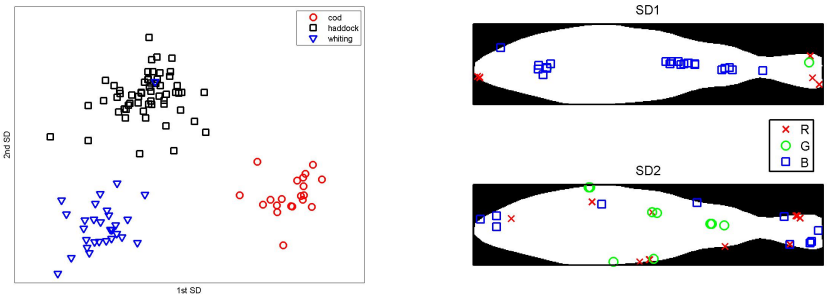


Figure C.8: Sparse discriminant directions in SDA. To the left, the clusters of three fish species. To the right, the selected texture features are marked on the fish mask. The first SD is mainly dominated by blue intensities whereas the 2nd SD consists of both red, green, and blue intensities. Only texture features were selected in the SDA.

have proposed algorithms for computing sparse versions of linear discriminant analysis and mixture discriminant analysis. The methods are especially useful when the number of observations is small in relation to the number of variables ($n \ll p$). Also they are generally useful when it is important to gain knowledge of a subset of features which separates two or more groups in high-dimensional problems. Sparse discriminant analysis has been illustrated on a small shape based data set of female and male silhouettes, a benchmark microarray data set for classification of leukemia subtypes, on visual and chemical data for classification of the fungi to the species or the genus level, and on shape and texture features for classification of fish species. Sparse mixture discriminant analysis was illustrated on the microarray data and the chemical data for classification of fungi to the genus level. The methods are in general faster than the methods which perform feature selection followed by a classification analysis, but not as fast as the shrunken regularized discriminant analysis algorithm. Furthermore, the classification results are comparable or even better than those obtained from the standard methods as well as for sparse partial least squares (SPLS) and shrunken centroids regularized discriminant analysis (RDA). In general, the methods outperformed SPLS in classification rates whereas they gave comparable classification results to those for RDA. However, RDA is conservative with regards to feature selection and therefore the methods proposed here provide sparser models and additionally provide low dimensional views of the data. It should also be noted that RDA is sensitive to normalization of data in terms of classification rates. Finally, the mixture of Gaussians models are useful for modelling data where biological subgroups exist such as classification of biological data to the species or the genus level. Matlab and R versions of SDA and

SMDA are available from: `www.imm.dtu.dk/~lhc`.

Acknowledgements

The authors would like to thank Gritt Rasmussen, Thomas Ostenfeld Larsen, Charlotte Held Gotfredsen and Michael E. Hansen at BioCentrum, The Technical University of Denmark for making the 1D NMR data available. Also thanks to Hildur Ólafsdóttir and Rasmus Larsen at Informatics and Mathematical Modelling, Technical University of Denmark for making the silhouette and fish data available, and to Karl Sjöstrand for valuable comments. Finally, the authors would like to thank the editor, an assoicate editor and two referees for valuable comments.

C.7 Appendix

C.7.1 The relation between optimal scoring and discriminant analysis

It is convenient to make the relation between the sparse optimal scoring criterion (C.8) and the sparse discriminant criterion (C.3) via canonical correlation analysis (CCA).

C.7.1.1 Sparse optimal scoring

The sparse optimal criterion in (C.8) is stated in terms of a single solution (θ, β), but implicitly it is a sequence of solutions (θ_j, β_j) with orthogonality given by the inner product $n^{-1} < Y\theta_j, Y\theta_l >= \delta_{jl}$ implied in the constraint (C.9). The sparse optimal scoring criterion can be rewritten to

$$ASR(\theta_j, \beta_j) = \theta_j^T \Sigma_{11} \theta_j - 2\theta_j^T \Sigma_{12} \beta_j + \beta_j^T \Sigma_{22} \beta_j + \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \quad , \qquad \text{(C.30)}$$

which is to be minimized under the constraint

$$\theta_j^T \Sigma_{11} \theta_j = 1 \quad , \qquad \text{(C.31)}$$

and where

$$\Sigma_{11} = n^{-1}Y^TY \tag{C.32}$$
$$\Sigma_{22} = n^{-1}(X^TX + \lambda_2\Omega) \tag{C.33}$$
$$\Sigma_{12} = n^{-1}Y^TX \quad ; \quad \Sigma_{21} = \Sigma_{12}^T \quad . \tag{C.34}$$

C.7.1.2 Sparse canonical correlation analysis

The sparse canonical correlation problem is defined by the criterion (which apart from the ℓ_1-term is the same as the penalized correlation problem, Hastie et al. (1995a))

$$COR_{\ell_1}(\theta_j, \beta_j) = \theta_j^T\Sigma_{12}\beta_j - \lambda_1\sum_{i=1}^{p}|\beta_{ji}| \quad , \tag{C.35}$$

which is to be maximized under the constraints

$$\theta_j^T\Sigma_{11}\theta_j = 1 \quad \text{and} \quad \beta_j^T\Sigma_{22}\beta_j = 1 \quad . \tag{C.36}$$

Under the CCA constraints we obtain $ASR = 2 - 2COR_{\ell_1}$, and the problems only differ in the additional constraint $\beta^T\Sigma_{22}\beta = 1$. Hence, for fixed θ the parameters in the optimal scoring problem β_{os} is, up to a scalar, the same as the parameters for the canonical correlation problem:

$$\beta_{j,cca} = \beta_{j,os}/\sqrt{\beta_{j,os}^T\Sigma_{22}\beta_{j,os}} \quad , \tag{C.37}$$

and the ℓ_1-weights are related as $\lambda_{1,cca} = \lambda_{1,os}/2$. Finally, we see that the optimal scores are the same for the two problems as we for fixed β have:

$$\theta_{cca} = \theta_{os} = \Sigma_{11}^{-1/2}UV^T \quad , \tag{C.38}$$

where $\Sigma_{11}^{-1}\Sigma_{12}\beta_{os} = US_{os}V^T$ or $\Sigma_{12}\beta_{cca} = US_{cca}V^T$.

C.7.1.3 Sparse discriminant analysis

The sparse discriminant analysis is defined as in (C.3)

$$BVAR_{\ell_1}(\beta_j) = \beta_j^T\Sigma_B\beta_j - \lambda_1\sum_{i=1}^{p}|\beta_{ji}| \quad , \tag{C.39}$$

which is to be maximized under the constraint

$$WVAR(\beta_j) = \beta_j^T\Sigma_{W_p}\beta_j = 1 \quad , \tag{C.40}$$

and where

$$\Sigma_B = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \quad (C.41)$$

$$\Sigma_{W_p} = \Sigma_W + \lambda_2 n^{-1}\Omega = \Sigma_{22} - \Sigma_B \quad . \quad (C.42)$$

Recall from penalized discriminant analysis (Hastie et al. (1995a)) that without the ℓ_1-penalization then the penalized discriminant analysis and penalized canonical correlation analysis coordinates are related as

$$\beta_{j,lda} = \beta_{j,cca} / \sqrt{\beta_{j,cca}^T \Sigma_{W_p} \beta_{j,cca}} \quad . \quad (C.43)$$

Comparing $BVAR_{\ell_1}$ (C.39) and COR_{ℓ_1} (C.35) and keeping in mind that the constraints are the same as under PDA it is easy to see that the relation still holds, and that the ℓ_1-weights are related as $\lambda_{1,lda} = \lambda_{1,cca}$.

C.7.1.4 Optimal scoring and discriminant analysis

Finally, we have the relation between sparse discriminant analysis and sparse optimal scoring given via their relations to CCA:

$$\beta_{lda} = \beta_{os} / \sqrt{\beta_{os}^T \Sigma_{W_p} \beta_{os}} \quad . \quad (C.44)$$

Furthermore, the ℓ_1-weights are related as $\lambda_{1,lda} = \lambda_{1,os}/2$.

Algorithm 2 Sparse Mixture Discriminant Analysis:

1. Initialize the blurred response matrix Z with the subclass probabilities. As in Hastie and Tibshrani (1996) the subclass probabilities can be derived from Learning Vector Quantization or K-means preprocessing, or from a priori knowledge of data. Initialize $\theta = (R\sum_{j=1}^{k}\sum_{r=1}^{R_j}\pi_{jr})^{-1}I_{1:R-1}$.

2. For $j = 1, ..., Q$, $Q \leq R-1$ solve the modified elastic net problem with fixed θ

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|Z\theta_j - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \qquad \text{(C.24)}$$

3. For fixed β and $Y^T\hat{Y} = USV^T$ compute the optimal scores

$$\theta = D_p^{-\frac{1}{2}}UV^T \quad , \qquad \text{(C.25)}$$

where D_p is a diagonal matrix of subclass probabilities, π_{jr}. π_{jr} is the sum of the elements in the r^{th} column in Z divided by the number of samples n.

5. Update the subclass probabilities in Z and the mixing probabilities in D_p using the estimation steps (C.20) and (C.21).

6. Repeat step 2-5 until convergence or a maximum number of iterations is reached.

7. Remove the last $R-m$ trivial directions, where the $(m+1)^{th}$ singular value $S_{m+1} < \epsilon$ (ϵ is some small threshold value):

$$\theta = D_p^{-\frac{1}{2}}UV_{1:m}^T \quad , \qquad \text{(C.26)}$$

For $j = 1, ..., m$ solve the modified elastic net problem with fixed θ using (C.24) to obtain the m nontrivial discriminant directions.

Appendix D

Classification of paired ear canal impressions in high dimensions - Data driven constraints for the Support Vector Machine

---

Authors: Line H. Clemmensen[1] and Sune Darkner[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

Submitted to *Medical Image Analysis*.

D.1 Abstract

Classification in longitudinal studies holds knowledge of the pairing of observations often unexploited when modeling the development of a specific condition. We propose a novel technique for classification between paired observations in high dimensions. The problem at hand is to classify whether ear canal impressions have been obtained with open or closed mouth. Studies have shown that the ear canal changes shape with movement of the mandible or head, affecting the physical comfort of the hearing aid for the user. Only $10-20\%$ exhibit deformation creating comfort problems and positive identification of these impressions are important. The proposed method achieves a correct classification rate of $60-70\%$, unachievable by audiologists.

The proposed method embeds knowledge of the pairing between samples to the support vector machine (SVM). A constraint is added to the SVM which strives for orthogonality of the vectors between paired observations and the estimated hyperplane in order to obtain more robust solutions. The primal and dual optimization problems are derived and extended to kernel space for the general case of adding data specific constraints to the SVM, and the usage of kernels under the ℓ_1-norm approximation is discussed. We show that the constrained SVM under the ℓ_2-norm generalizes to a linear kernel. Imposing the constraint of orthogonality on the paired data is more robust, compared to the SVM, with reduced variance and the classification rates are additionally improved. The method is tested on simulated data and a high-dimensional paired data set of ear canal surfaces.

D.2 Introduction

This paper examines shape differences of ear canals. Ear canals deform slightly with the movement of the mandible or turning of the head. The deformation occurs when chewing or moving the head. In some cases this leads to physical discomfort for hearing aid users. The data set considered in this paper consists of impressions from 67 individuals. Two impression have been acquired from each; one with open mouth and one with closed mouth. An additional impression was obtained from 42 of the subjects with their head turned.

When receiving impressions of ears for custom hearing aid production, it is not

known wether an impression has been obtained from an individual with open or closed mouth. It is desirable to be able to identify impressions of ears with a large deformation which have been acquired with the mouth open or preferably classify the shapes into open and closed mouth impressions. Since the variation between the individual ears is much larger than the deformation separating two impressions from the same individual the chance of a high overall classification rate is unlikely. Audiologists are in general not capable of classifying a single impression as taken with open or closed mouth. Previous, (Darkner et al., 2007) showed that all impression pairs exhibit a significant change between them from open to closed mouth. However, that does not reveal a general tendency in the deformations. Additionally, only 10-20% experience a deformation with a magnitude high enough to cause comfort problems. Our hope is that identification of potentially problematic cases will help in making more comfortable hearing aids and reduce the number of custom hearing aids that are remodeled. The classification problem can be solved using the support vector machine (SVM) in its normal formulation, however, we have more information at our disposal, i.e. the pairing of the observations which we would like to incorporate in the model to make it more robust and general.

Like in this case of pairing between observations, many data problems hold more information than what seems apparent. The structure of the data or the way the data is collected often contains information of covariances in feature space or a more direct linkage between observations, as the pairing of observations of the ears. Therefore, a constraint which can embed such knowledge into the SVM in general is proposed. Specifically, for paired observations a constraint is presented that enforces orthogonality of the vectors spanned by paired observations and the estimated hyperplane. The general framework for adding constraints based directly on the data to the SVM in the primal formulation is presented. The dual formulation is derived from the generalized primal formulation with constraints and hereby both classification and regression are extended to a non-linear kernel space. It is shown for a general constraint based directly on data that when selecting the appropriate formulation the constraint can be formulated as inner products. This means that kernels can be applied to expand the solution space to various more complex spaces such as e.g. polynomials, but without making an explicit feature repersentation (M. Aizerman and Rozonoer, 1964).

Recently, the SVM was extended to use an ℓ_1-norm instead of an ℓ_2-norm (Li et al., 2006), and to include both norms as in the elastic net formulation (Wang et al., 2006). The additional constraints added in this paper are based directly on data and therefore introduce additional information to the previous methods. On top of that the problem can be formulated using inner products and it is therefore possible to apply kernels and thus obtain non-linear classifiers. This is also known as the kernel trick which means that the input space as a function is expanded to higher dimensions and the classes are separated with a linear

hyperplane in the expanded space, yielding a non-linear separation when projected back into the input space. In particular, when observations are paired as in certain types of medical studies, we consider adding a constraint that enforces orthogonality between the separating hyperplane and the vectors spanned by the paired observations in two groups. Adding this constraint reduces the variance of the model by introducing a bias. It will be shown that when such a pairing of data exists the bias/variance trade off ensures more robust results while the classification error is statistically comparable or slightly better compared to that of ordinary SVM.

For the ear canal problem at hand it is, apart from the classification task, also of interest to find the direction which describes the difference between the two classes. This direction is given implicitly by the SVM framework as the normal vector to the separating hyperplane, and has been described in detail in (Golland, 2001).

In the following some of the previous work on the support vector machine will briefly be summarized, the addition of the data driven constraints to the SVM will be described, and its performance on simulated data as well as real data for the orthogonality constraint on paired data will be illustrated. Finally, the paper is summed up with a discussion.

D.3 Summary of previous work on the SVM

The support vector machine (SVM) was introduced in Boser et al. (1992). The SVM builds on theory for the optimal separating hyperplane developed by Vapnik and Chervonenkis in 1965 (Vapnik, 1999) but constructs the hyperplane in a kernel feature space, and therefore the feature space does not have to be in explicit form. SVMs are most known for their use in classification problems but other uses such as regression have been proposed (Shawe-Taylor and Cristianini, 2004; Hastie et al., 2009). The SVM is in its simplest form defined as the separating hyperplane given by:

$$\boldsymbol{y} = \boldsymbol{\beta}^t \mathbf{x} + \boldsymbol{\beta}_0 \tag{D.1}$$

where t denotes the transpose and y_i is coded according to the class of $\mathbf{x}_i$: $\{-1, 1\}$. The SVM maximizes the distance between the observations defining the boundary of the two classes. Since the groups are not always separable, slack variables ξ_i are introduced to allow for misclassification. Figure D.1 illustrates the SVM. We use the following formulation of the minimization problem

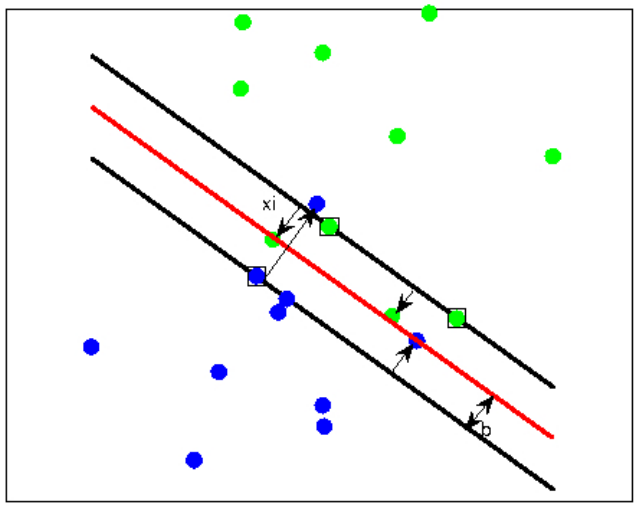


Figure D.1: Geometry of the SVM. The decision boundary separates the two classes and the margin of total width $\frac{2}{\|\beta\|}$ is created. Each point inside the margin is associated with a slack variable ξ_i, the single arrows illustrate the slack variables. The margin is maximized while keeping the sum of the slack variables under some constant. The points marked with squares are support vectors which lie on the margin.

$$\begin{aligned} &\min \|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i \\ &\text{subject} \quad \text{to}: \quad y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \quad , \end{aligned} \tag{D.2}$$

where γ is a constant determining the weight on the slack variables (the misclassifications and points inside the margin) in relation to the weight on the margin size. In the case where the number of variables p is the same or larger than the number of observations n the SVM suffers from instability issues due to the curse of dimensionality (Hastie et al., 2009). However, by reformulating the problem to its Lagrange dual (Vapnik, 1999) the solution can be found in the n-dimensional space like in ridge regression (Hoerl and Kennard, 1970) to which the regression problem can be easily reformulated (Shawe-Taylor and Cristianini, 2004; Hastie et al., 2009). The SVM for regression has been modified earlier to embrace the LASSO constraint for regression (Li et al., 2006) but has in this form not been generalized to classification, different loss functions of the residual (Hastie et al., 2009), or the doubly regularized SVM as a classification

version of the elastic net regression (Wang et al., 2006; Zou and Hastie, 2005).

Finally, Golland (2001) originally defined a discriminative direction to be the direction which moves a point towards the other class while introducing as little irrelevant change as possible with respect to the classifier function. This idea is used to visualize the discriminative direction in the original space of the shape models and additionally to use the physical properties of data to obtain a variance reduction in the classifier function.

D.4 Methodology

This section derives the quadratic optimization problem used for classification with added prior information such as pairing of observations. Section D.4.1 adds a regularization term to the SVM using an ℓ_1-norm approximation. The primal and dual formulations of the SVM are derived to show how and when kernels can be applied in the formulation. The third section discusses the added constraint in general whereas section D.6 introduces the constraint of orthogonality used for data with paired observations.

The ℓ_2-norm gives more weight to outliers than the ℓ_1-norm and therefore has not been considered in the experiments here, but the derivation under the ℓ_2-norm penalty is added in appendix.

D.4.1 ℓ_1-norm constraint

The ℓ_1-norm does not punish outliers as heavily as the ℓ_2-norm and in many cases when the constraint depends on data and we want to regularize according to the trend and not according to outliers it is a more appropriate choice than the ℓ_2-norm. The SVM is formulated with an ℓ_1-norm of $\boldsymbol{\beta}\mathbf{A}$ added $\|\boldsymbol{\beta}^t\mathbf{A}\|_1$, where $\mathbf{A}$ is a $p \times m$ matrix. However, since the ℓ_1-norm is not differentiable at zero the problem must be modified to avoid this singularity. Here, the ℓ_1 is approximated by letting $\boldsymbol{\delta} \geq \boldsymbol{\beta}\mathbf{A} \geq -\boldsymbol{\delta}$ and then minimizing with respect to δ. This approximation furthermore gives a suitable formulation of the dual problem and ensures, as we see later, that the kernel trick can be applied. The

primal SVM problem with the added constraint is

$$\min \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i + \lambda \sum_{k=1}^{m} \delta_k \tag{D.3}$$
$$\text{s.t.} \begin{cases} -\delta_k \leq \boldsymbol{\beta}^t \mathbf{a}_k, \quad \delta_k \geq \boldsymbol{\beta}^t \mathbf{a}_k \quad \forall k \\ y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \end{cases}$$

The Lagrange primal can then be written as

$$\begin{aligned} L_P &= \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i + \lambda \sum_{k=1}^{m} \delta_k - \sum_{k=1}^{m} \rho_k(\boldsymbol{\beta}^t \mathbf{a}_k + \delta_k) + \sum_{k=1}^{m} \rho'_k(\boldsymbol{\beta}^t \mathbf{a}_k - \delta_k) \\ &\quad - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{n} \mu_i \xi_i \end{aligned} \tag{D.4}$$

Differentiating L_P with respect to $\boldsymbol{\beta}$, β_0, ξ_i, δ_k and δ'_k and equating to zero, the following is obtained

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i + \sum_{k=1}^{m} \rho_k \mathbf{a}_k - \sum_{k=1}^{m} \rho'_k \mathbf{a}_k \tag{D.5}$$
$$0 = \sum_{i=1}^{n} \alpha_i y_i \tag{D.6}$$
$$\alpha_i = \gamma - \mu_i \tag{D.7}$$
$$0 = \lambda - \rho_k - \rho'_k \tag{D.8}$$

as well as the positive constraints $\alpha_i, \mu_i, \xi_i \geq 0 \ \forall i$ and $\rho_k, \rho'_k, \delta_k \geq 0 \ \forall k$. By insertion of by (D.6),(D.7) and (D.8) in (D.4) we get the dual objective function

$$\begin{aligned} L_D &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j - \sum_{k=1}^{m} \sum_{i=1}^{n} \alpha_i y_i (\rho'_k - \rho_k) \mathbf{x}_i^t \mathbf{a}_k \\ &\quad + \frac{1}{2} \sum_{k=1}^{m} \sum_{l=1}^{m} (\rho'_k - \rho_k)(\rho'_l - \rho_l) \mathbf{a}_k^t \mathbf{a}_l \end{aligned} \tag{D.9}$$

Maximizing L_D subject to $0 \leq \alpha_i \leq \gamma$, $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\rho'_k + \rho_k = \lambda$ yield the desired result. Additional to (D.5)-(D.8) the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939) are given as:

$$\rho_k(\boldsymbol{\beta}^t \mathbf{a}_k + \delta_k) = 0, \quad \rho'_k(\boldsymbol{\beta}^t \mathbf{a}_k - \delta_k) = 0 \quad \forall k \tag{D.10}$$
$$\alpha_i[y_i(\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) - (1 - \xi_i)] = 0 \tag{D.11}$$
$$\mu_i \xi_i = 0 \tag{D.12}$$
$$y_i(\mathbf{x}_i^t \boldsymbol{\beta}) - (1 - \xi_i) \geq 0 \quad . \tag{D.13}$$

From (D.11) it is seen that the points on the margin ($\alpha_i = 0$) can be used to calculate β_0. The hyperplane is written as

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}^t \mathbf{x}_i + \sum_{k=1}^{m} (\rho'_k - \rho_k) \mathbf{x}^t \mathbf{a}_k \quad . \tag{D.14}$$

If $\mathbf{A}$ is defined as a function of the input space $\mathbf{x}$, that is $\mathbf{A} = \mathbf{XB}$, where $\mathbf{X}$ is a matrix representing all observations then $\mathbf{x}^t\mathbf{a}_k$ is an inner product. Then, since kernels are applicable to all inner products of the input space $\mathbf{x}$, kernels can be applied such that non-linear separations can be obtained (also known as the *kernel trick*). If this is not the case, the feature space of the kernel $\phi(\mathbf{x})$ must be explicitly known. When the constraint on $\boldsymbol{\beta}$ is not a function of $\mathbf{x}$ it is desirable to know the feature space explicitly in order to add a meaningful constraint. The choice of $\mathbf{A}$ will be addressed in the following sections.

A special case is when $\mathbf{A} = \mathbf{I}$; this corresponds to an elastic net formulation of Zou and Hastie (2005); Wang et al. (2006) where both the Ridge (ℓ_2-norm Hoerl and Kennard (1970)) and the LASSO (ℓ_1-norm Tibshirani (1996)) constraints are added.

D.4.2 ℓ_2-norm constraint

The addition of the constraint on $\boldsymbol{\beta}$ under the ℓ_2-norm $\|\boldsymbol{\beta}^t\mathbf{A}\|_2^2$ is derived. This can be formulated as the following minimization problem

$$\begin{aligned} &\min \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\|\boldsymbol{\beta}^t\mathbf{A}\|_2^2 \\ &\text{s.t.} \quad y_i(\mathbf{x}_i^t\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \end{aligned} \tag{D.15}$$

where γ is a constant weight on the slack variables and λ is a weight on the introduced constraint. The primal Lagrange function can then be stated as

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{x}_i^t\boldsymbol{\beta} + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{n} \mu_i \xi_i + \frac{\lambda}{2}\|\boldsymbol{\beta}^t\mathbf{A}\|_2^2 \tag{D.16}$$

By differentiating with respect to $\boldsymbol{\beta}$, β_0 and ξ_i the following expressions are obtained.

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} \mathbf{x}_i \tag{D.17}$$

$$0 = \sum_{i=1}^{n} \alpha_i \mu_i \tag{D.18}$$

$$\alpha_i = \gamma - \mu_i \tag{D.19}$$

as well as the positive constraints $\alpha_i, \mu_i, \xi_i \geq 0 \ \forall i$. By insertion in (D.16) we get the dual objective function (c.f. App. D.9)

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} \mathbf{x}_j \tag{D.20}$$

which we maximize subject to $0 < \alpha_i < \gamma$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$, and the KKT conditions include the constraints

$$\alpha_i [y_i (\mathbf{x}_i^t \boldsymbol{\beta} + \beta_0) - (1 - \xi_i)] = 0 \tag{D.21}$$

$$\mu_i \xi_i = 0 \tag{D.22}$$

$$y_i (\mathbf{x}_i^t \boldsymbol{\beta}) - (1 - \xi_i) \geq 0 \tag{D.23}$$

We see that the matrix $(\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1}$ in (D.20) simply is a linear kernel applied to the input space x. It is possible to apply another kernel if this kernel is linear since it then holds that $< \phi(\Phi(x_i)), \phi(\Phi(x_j)) > = < \Phi(\phi(x_i)), \Phi(\phi(x_j)) >$, where $\Phi(.)$ and $\phi(.)$ are some linear kernels, thus (D.20) can then be formulated as

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K((\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1/2} \mathbf{x}_i, (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1/2} \mathbf{x}_j) \quad . \tag{D.24}$$

A Taylor expansion of a nonlinear kernel can be used to obtain linearity.

D.5 Constraints for paired observations

Consider data with two classes of paired observations also called matched points, i.e. an observation in one class has a natural pairing with an observation in the second class and they are therefore not independent. This is for example the case when the same individual is observed at two stages or when subjects are matched according to some criterion such as age or family background. If we want to separate the two classes with a traditional clustering method or

classifier (Duda et al., 2001; Hastie et al., 2009) the information about the pairing is not exploited. In particular when $p \gg n$, exploiting such information is crucial in order to avoid the curse of dimensionality (Hastie et al., 2009). Specifically for the orthogonality constraint (OC) we set $\mathbf{A} = (\mathbf{X_1} - \mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are matched points sets, the following is obtained for the ℓ_2-norm

$$\mathbf{A}^t\mathbf{A} = (\mathbf{X_1} - \mathbf{X_2})^t(\mathbf{X_1} - \mathbf{X_2}) = \mathbf{X_1}^t\mathbf{X_1} + \mathbf{X_2}^t\mathbf{X_2} - \mathbf{X_1}^t\mathbf{X_2} - \mathbf{X_2}^t\mathbf{X_1} \quad \text{(D.25)}$$

where $\mathbf{X_1}$ contains observations belonging to group 1 and $\mathbf{X_2}$ contains the ordered counterpart belonging to group 2, i.e. $\mathbf{a}_k = \mathbf{x_{1}}_k - \mathbf{x_{2}}_k$.

The hyperplane for the OC-SVM under the ℓ_1-norm is then given by

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}^t\mathbf{x}_i + \sum_{k=1}^{n/2} (\rho'_k - \rho_k)\mathbf{x}^t(\mathbf{x_{1}}_k - \mathbf{x_{2}}_k) \quad . \quad \text{(D.26)}$$

The optimization problem is expressed in terms of inner products in the input space, and therefore, in principle, the kernel trick can be applied.

The cosine of the standard inner product emphasizes large angles in this setting and thereby gives weight to outliers, i.e. directions which are far from orthogonal to the separating hyperplane. It is not of interest to emphasize outliers further by using the ℓ_2-norm and the kernel generated for ℓ_2-norm have dimension $p \times p$, which in practice is impossible for problems with more than 10^4 variables due to memory restrictions. It is of interest to get a more general picture of the difference and therefore the ℓ_1-norm is preferable. Another factor is the length of $\mathbf{a}$. Since $|\boldsymbol{\beta}|$ is constant the length of $\mathbf{a}$ acts like a weight where large differences in terms of distance between observations are weighted proportionally to this distance. To circumvent this, the directional vectors of $\mathbf{A}$ could be normalized. However, as only individuals with large deformations in the ear canals are actually affected with regards to discomforts wearing hearing aids it is of interest to put a higher emphasis on exactly these individuals and thus we do not normalize here.

It can be useful to solve the dual problem even without use of kernels as the dual problem is solved in the observation space. Hence, for problems with $p \gg n$ the dual problem has much smaller dimensions than the corresponding primal problem.

D.6 General constraints

In general the matrix $\mathbf{A}$ can be any matrix, it can form the derivatives of $\boldsymbol{\beta}$ like in Fused Lasso (Tibshirani and Saunders, 2005), it can perform feature selection

with $\mathbf{A} = \mathbf{I}$ under the ℓ_1-norm (Zou and Hastie, 2005) or it can be dependent on data. However, when $\mathbf{A}$ is independent of data the kernel trick can not be applied.

When the matrix $\mathbf{A}$ is dependent on the data x then $\mathbf{A}^t\mathbf{A}$ can be written as purely inner products of the data and therefore the kernel trick is in principle applicable, i.e. $\mathbf{A} = \mathbf{XB}$, where $\mathbf{B}$ is some $n \times m$ matrix. However, it is important to consider whether the introduced constraint is meaningful in kernel space. In the next section we will look at the constraint of orthogonality for paired observations.

D.7 Experiments

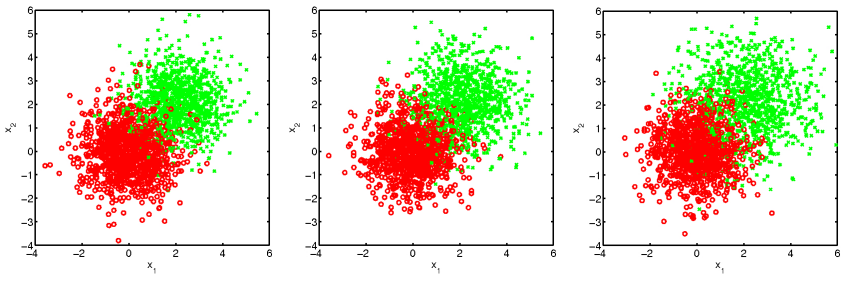
The performance of the OC-SVM is tested on synthetic as well as real data and compared to the performance of the standard SVM. The synthetic data serves to visualize the performance of OC-SVM as well as to test the performance for various numbers of observations and dimensions. Finally the methods are tested on real high dimensional data in terms of 172 ear canal impression scannings from 67 individuals.

D.7.1 Synthetic data

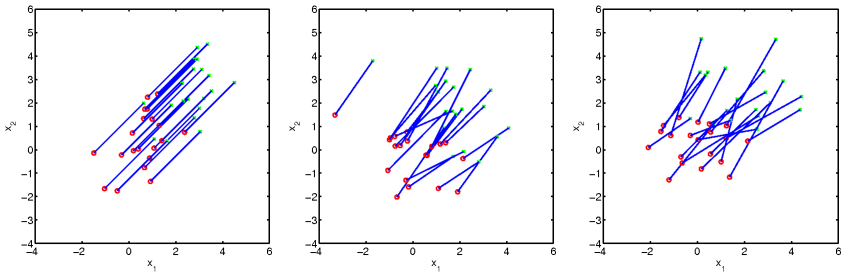
The synthetic data was generated from a normal distribution with standard deviation one (group one). To simulate the coherence between the two groups all points in group one were translated by $\sqrt{\frac{1}{p}}$, where p is the dimension of the feature space, along each axis in the positive direction. White noise with standard deviation, $\sigma = 0.0$, 0.6 and 0.9 was added to each of the translated data points (group two). This ensured that the two distributions were non-separable and had approx. 70% overlap. Figure D.2 shows the distribution of the two generated classes in 2D for 1000 data points in each class and with no noise added, with noise of standard deviation 0.6, and noise of standard deviation 0.9.

To illustrate OC-SVM the separating lines for SVM and OC-SVM are plotted in two dimensions; see figure D.3. It is seen that OC-SVM reduces the variance of the solution, but introduces a small bias, i.e. the hyperplanes shift parallel, but rotate less for OC-SVM.

Using this data, 1000 paired samples were generated from which the training



(a) $\sigma = 0$ (b) $\sigma = 0.6$ (c) $\sigma = 0.9$



(d) $\sigma = 0$ (e) $\sigma = 0.6$ (f) $\sigma = 0.9$

Figure D.2: The synthetic data with added noise to the translation from group one to two group with $\sigma = 0$, $\sigma = 0.6$ and $\sigma = 0.9$, and figures of 40 samples where the vector between each paired samples are illustrated.

and test sets were drawn. To properly test the method the regularization parameters were changed in such a way that the standard SVM estimates the optimal separating hyperplane including all samples. The purpose was to show that the added constraint actually influenced the result. The OC-SVM regularization parameter was fixed at 10% of the SVM regularization, which from our experience is a good heuristic. Samples were generated in 2, 5, 10, 100, 500, 1000 and 10000 dimensions and the models were for each dimension built on 2, 5, 10, 50 and 100 paired samples. For each combination the experiment was repeated 100 times where the SVM and the OC-SVM were built on the same samples. The training error and the test error were measured for each method and a hypothesis of the test error being equal for the two methods was tested. The resulting p-values for the hypothesis testing and the misclassification rates for the two methods are summarized in tables D.1-D.3 in appendix D.10.

The solutions of the OC-SVM have far less variance than the normal SVM, thus the variance over 100 solutions is as little as some 80% of the variations of the solutions from the SVM. The simulations show that the added information

in general gives much more stability to the OC-SVM compared to the SVM. However, when the noise of the pairing is large, here $\sigma = 0.9$ it is seen that gain in stability is smaller. This is to be expected when the information of the pairing is noisy, i.e. the additional constraint builds on information with a low signal to noise ratio. Furthermore, the classification results of OC-SVM in general gets poorer as the paired information gets noisier.

In the simulation study it is seen that OC-SVM gives significantly lower misclassification rates when the number of observations is of the same size as the number of variables or down to around 1/10 of the number of variables. This trend is also seen when noise is added. When the number of observations is either larger than, or much smaller than the number of variables, no significant difference between the two methods is observed.

The extra bias introduced marginally improves performance in some cases and does not decrease performance in general (note, that the weight of the pairing constraint can be set to zero and the two solutions would be the same). In addition to this it is obvious that the OC-SVM significantly reduces the amount of variation between solutions built from the same distribution. In the next section this variance reduction effect becomes even more pronounced as we study the ear experiment.

D.7.2 Ear data

This data is part of a larger study of how the ear canal changes shape due to movement of the mandible, turning of the head and leaning over. The purpose of the study is to examine and quantify the changes and survey the resulting influences on the comfort for the hearing aid user. For all individuals the shape of the ear canal changes when the mouth is opened or the head is turned, but for some hearing aid users this causes discomfort either as direct physical discomfort or in the form of acoustical feedback due to what is known as a false vent. This means that the acoustic properties to which the hearing aid is adjusted have changed sufficiently to cause an acoustical feedback loop. In the hearing aid industry the hearing aids are sold to the consumer through an audiologist. Their practice with respect to acquisition of the impressions influences the shape and thus the comfort level of the individual hearing aid. There is no standardized way of obtaining impressions, different audiologist have different practices: Some are using open ear impressions and some closed. If a hearing aid is rejected by the user the hearing aid manufacturer takes the economical loss of the returned hearing aid, which in the end increases the overall cost of hearing aids. Therefore, the manufacturer has an interest in being able to classify a given impression into whether it was taken with open or closed mouth. This is

in particular of importance if the ear exhibits a large change due to jaw movement, information which is not passed along with the impression. The hope is that eventually an online classification of the shape will be enabled along with guidelines for best practice in this respect.
The impressions were obtained from each individual's right ear; one with closed mouth, one with open mouth and one with the head turned to the left. The impressions were made by the same audiologist to ensure consistency and scanned by the same operator on a 3D-laser scanner. Figure D.6 shows the resulting surface scanning of an ear impression. The data analyzed consists of 42 triplet pairs of impressions, a total of 126 impressions for the three settings: open mouth, closed mouth and turning the head, plus an additional 25 pairs for the open and closed mouth (67 pairs in total for open and closed mouth setting).

All impressions have been registered to create a common frame of reference, see Darkner et al. (2007); resulting in 4356 points in 3D with full correspondence. Each individual ear is represented as a 13068 dimensional vector of the these coordinates.

We permute the order of the observations 500 times and each time train on the first $n-2$ impressions and test on the two last, one from each class and thereby obtain measures of classification rates, variance in the model and significance of improvement. This is done for classification between open mouth and closed mouth and between closed mouth and turning of the head. For one of the data sets the CPU time was computed to 0.03s and 0.18s for SVM and OC-SVM, respectively.

Between open and closed mouth, the average training errors were 0% for SVM and 38% for OC-SVM. Both SVM and OC-SVM had average test errors of 40%. We see that SVM overfits the training data too a much higher degree than OC-SVM. The most important result was that none of the large deformations capable of causing physical discomfort were misclassified. For the majority of the misclassification the deformations were small, i.e. around and below average which is approximately 0.7 mm for which the incompliance can be absorbed by the soft tissue of the ear which maintains the acoustical seal.

Since the method regularizes the model and thereby increases the bias, the variance should be reduced and a more generalizable result obtained. Comparing the average variance of the results obtained with the OC-SVM and SVM respectively shows that the additional constraint reduced the variance of the normalized $\boldsymbol{\beta}$ with 50%. The overall average variances were $1.1642 \cdot 10^{-5}$ and $6.9966 \cdot 10^{-6}$ for SVM and OC-SVM, respectively. The variation of $\boldsymbol{\beta}$ over the 500 permutations can be seen in Figure D.7. Furthermore the variance of the estimates of the intercept for the OC-SVM is 100 times lower than the SVM and the variation in the length of the parameter vector is 3000 times lower than

that of the SVM.

Additionally, we applied the OC-SVM to distinguish between closed mouth impressions and turning of the head impressions. The results showed that the SVM had an average correct test classification of 67% and the OC-SVM around the same. However, the formulation provided above allows us to use kernels. When we applied a polynomial kernel of degree 2 we increased the average test classification success with an additional 3% to roughly 70% for the OC-SVM only.

We also tried to apply the OC-constraint under the ℓ_2-norm. In order to be able to form and invert the linear kernel the constraint reduces to, the number of variables was reduced by a factor of 2. The method is not able to achieve classification better than random whereas the OC-SVM on the same data performed as reported above. The reason is most likely that the method put too much emphasis on outliers as mentioned in the previous section.

Figure D.8 illustrates the discriminative direction of a solution for OC-SVM. The general areas which change when the mouth is opened are the ear canal and the concha cymba. As the figure shows the discriminant areas are around the canal (left side of figure), however also the shape of Concha Cymba (left part) seems to have importance in the classification.

D.8 Conclusion

A classification model based on the SVM with additional constraints based on knowledge of data has been derived. For the constraint of orthogonality on paired data the variance of the separating hyperplane was reduced by up to 50% leading to more robust solutions. For the open closed mouth case the majority of the potential problematic cases where identified. For the turning of the head the kernel trick where applied with a decrease in classification error of 3% as a result. In both cases the performance of the OC-SVM where equally good or better compared to the SVM. This is of great importance, in particular when few observations are available and the variance in general is known to be high due to the curse of dimensionality. Furthermore, the classification rates for the OC-SVM proved to be significantly better or comparable to those of the ordinary SVM. For the OC-SVM the paired observations are automatically weighted according to the Euclidian length of the difference vector between the paired observations. That is, paired observations with large differences have a higher weight in the constraint of orthogonality than paired observations with subtle differences. In extension to that, a general framework for adding data

specific constraints to the SVM was derived in this paper. The framework makes it easy to use underlying a priori knowledge of data to obtain robust solutions for classification problems. On top of that the framework may also be used to obtain certain desired properties for the solutions such as sparseness or correlation between variables.

D.9 APPENDIX: ℓ_2-norm

In this section we give details on the derivation of the Lagrange dual, L_D for the ℓ_2-norm constraint. Differentiating (D.16) with respect to $\boldsymbol{\beta}$ and equating to zero gives

$$\begin{aligned} \boldsymbol{\beta} &= \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i - \lambda \mathbf{A}^t \mathbf{A} \boldsymbol{\beta} \Leftrightarrow (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A}) \boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \Leftrightarrow \\ \boldsymbol{\beta} &= (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \end{aligned} \tag{D.27}$$

By insertion of (D.17)-(D.19) in (D.16) and setting $X = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ we get the dual objective function

$$\begin{aligned} L_D &= \frac{1}{2} \|(\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X\|^2 + \sum_{i=1}^{n} \alpha_i + \frac{\lambda}{2} \|(\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X)^t \mathbf{A}\|^2 - X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X \\ &= \sum_{i=1}^{n} \alpha_i + \frac{1}{2} X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-t} (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X + \frac{\lambda}{2} X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-t} \mathbf{A}^t \mathbf{A} (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X \\ &\quad - X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X \\ &= \sum_{i=1}^{n} \alpha_i + \frac{1}{2} X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A}) (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X - X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} X^t (\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})^{-1} X \end{aligned}$$

since the matrix $(\mathbf{I} + \lambda \mathbf{A}^t \mathbf{A})$ is symmetric.

D.10 APPENDIX: Tables summarizing the results for the synthetic data experiments

A 1000 samples were generated in 2, 5, 10, 100, 500, 1000 and 10000 dimensions and the models were for each dimension built on 2, 5, 10, 50 and 100 paired samples, and the test error calculated on the remaining of the 1000 samples. For each combination the experiment was repeated 100 times where the SVM and the OC-SVM were built on the same subset of samples at each run. The training error and the test error were measured for each method and a hypothesis of the test error being equal for the two methods was tested. It is seen that OC-SVM

Table D.1: The p-values for the significance of the difference between the misclassification rates of OC-SVM and SVM using a paired t-test (H_0 = no difference) over 100 solutions. In parentheses the average misclassification rates of OC-SVM and SVM are given, the best classifier is in bold. The data are with added noise of the size $\sigma = 0$.

$p\backslash n$	2	5	10	50	100
10	0.0000 (**11.10** - 12.17)%	0.0000 (**9.28** - 13.19)%	0.0773 (**8.31** - 8.45)%	0.9715 (**7.89** - 7.89)%	0.4706 (**7.63** - 7.65)%
100	0.2999 (**11.16** - 11.25)%	0.0000 (**8.96** - 9.28)%	0.0000 (**7.77** - 8.74)%	0.0000 (**7.70** - 9.76)%	0.8289 (**7.72** - 7.75)%
500	0.0064 (**10.46** - 10.53)%	0.0003 (**8.99** - 9.11)%	0.0000 (**7.93** - 8.07)%	0.0000 (**7.02** - 7.91)%	0.0000 (**7.13** - 9.03)%
1000	0.0188 (**11.30** - 11.33)%	0.1282 (**8.29** - 8.32)%	0.0000 (**7.63** - 7.75)%	0.0000 (**6.95** - 7.39)%	0.0000 (**7.03** - 7.93)%
10000	0.9542 (**10.89** - 10.89)%	0.6209 (8.44 - **8.44**)%	0.2968 (**7.73** - 7.74)%	0.0569 (**6.94** - 6.97)%	0.0007 (**6.80** - 6.87)%

gives significantly lower misclassification rates when the number of observations is of the same size as the number of variables or down to around 1/10 of the number of variables. This trend is also seen when noise is added. When the number of observations is either larger than, or much smaller than the number of variables, no significant difference between the two methods is observed.

In tables D.4-D.6 the variances and relative differences in variance between OC-SVM and SVM are summarized. The solutions of the OC-SVM have far less variance than the normal SVM, thus the variance over 100 solutions is as little as 60% of the variations of the solutions from the SVM. In particular when the number of observations is small, the reduction in variance can be observed. The tables show that the added information in general gives much more stability to the OC-SVM compared to the SVM. However, when the noise of the pairing is large, here $\sigma = 0.9$ it is seen that gain in stability is smaller. This is to be expected when the information of the pairing is noisy, i.e. the additional constraint builds on information with a low signal to noise ratio.

Table D.2: The p-values for the significance of the difference between the misclassification rates of OC-SVM and SVM using a paired t-test (H_0 = no difference) over 100 solutions. In parentheses the average misclassification rates of OC-SVM and SVM are given, the best classifier is in bold. The data are with added noise of the size $\sigma = 0.6$.

$p \backslash n$	2	5	10	50	100
10	0.0000 (**16.09** - 18.10)%	0.0001 (**12.44** - 13.02)%	0.0538 (**10.71** - 10.92)%	0.7664 (**9.74** - 9.75)%	0.6943 (**10.55** - 10.56)%
100	0.7850 (36.32 - **36.28**)%	0.0009 (**23.15** - 23.81)%	0.0005 (**15.91** - 16.34)%	0.0765 (11.13 - **10.92**)%	0.0823 (**10.12** - 10.39)%
500	0.3571 (49.92 - **49.92**)%	0.0018 (47.42 - **47.32**)%	0.0483 (38.58 - **38.40**)%	0.0000 (16.10 - **15.56**)%	0.0000 (12.84 - **12.22**)%
1000	1.0000 (**50.00** - 50.00)%	0.3120 (49.86 - **49.86**)%	0.0000 (47.64 - **47.52**)%	0.1222 (**22.76** - 22.90)%	0.0000 (16.35 - **15.68**)%
10000	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	0.3197 (49.94 - **49.93**)%	0.0000 (47.84 - **47.76**)%

The extra bias introduced marginally improves performance in some cases and does not decrease performance in general. In addition to this it is obvious that the OC-SVM significantly reduces the amount of variation between solutions built from the same distribution. In the next section this variance reduction effect becomes even more pronounced as we study the ear experiment.

Table D.3: The p-values for the significance of the difference between the misclassification rates of OC-SVM and SVM using a paired t-test (H_0 = no difference) over 100 solutions. In parentheses the average misclassification rates of OC-SVM and SVM are given, the best classifier is in bold. The data are with added noise of the size $\sigma = 0.9$.

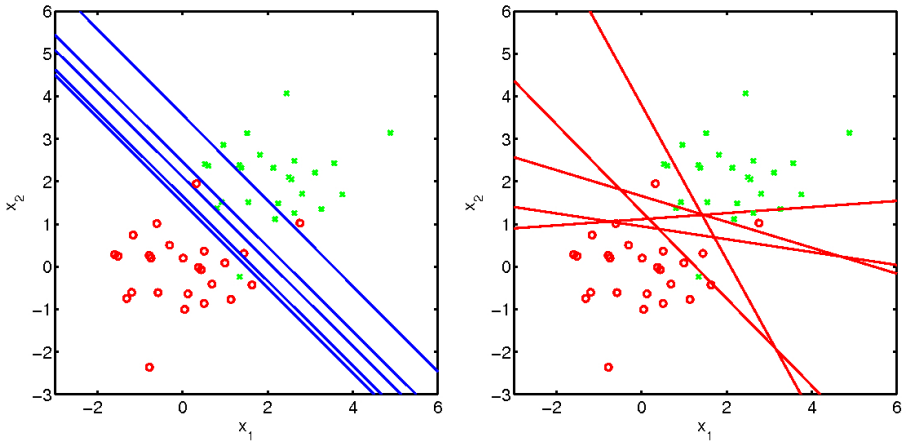
$p\backslash n$	2	5	10	50	100
10	0.0007 (**19.71** - 21.16)%	0.0591 (**16.32** - 16.65)%	0.1451 (**13.66** - 13.80)%	0.1592 (**12.85** - 12.95)%	0.1623 (**12.91** - 13.03)%
100	0.2187 (46.47 - **46.39**)%	0.1542 (34.55 - **34.33**)%	0.4111 (**24.71** - 24.85)%	0.2675 (15.53 - **15.34**)%	0.7943 (**14.08** - 14.13)%
500	0.3197 (50.00 - **50.00**)%	0.0004 (49.80 - **49.77**)%	0.0000 (47.44 - **47.29**)%	0.0000 (26.58 - **25.41**)%	0.4327 (20.37 - **20.15**)%
1000	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	0.5323 (49.82 - **49.82**)%	0.0000 (35.82 - **35.10**)%	0.0000 (26.85 - **25.44**)%
10000	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	1.0000 (**50.00** - 50.00)%	0.0003 (49.88 - **49.86**)%

Table D.4: The relative difference in average variance between OC-SVM and SVM, $\frac{Var(SVM)-Var(OC-SVM)}{Var(SVM)}$ of 100 solutions. In parentheses the average variances of OC-SVM and SVM are given. No noise added.

$p\backslash n$	2	5	10	50	100
10	83.45% (0.0020 - 0.0119)	78.79% (0.0066 - 0.0310)	37.50% (0.0004 - 0.0006)	7.82% (0.0000 - 0.0000)	36.12% (0.0000 - 0.0000)
100	86.88% (0.0000 - 0.0001)	86.43% (0.0001 - 0.0005)	85.95% (0.0001 - 0.0009)	79.35% (0.0003 - 0.0017)	11.86% (0.0000 - 0.0000)
500	87.18% (0.0000 - 0.0000)	87.11% (0.0000 - 0.0000)	86.99% (0.0000 - 0.0000)	85.98% (0.0000 - 0.0002)	84.54% (0.0001 - 0.0004)
1000	87.22% (0.0000 - 0.0000)	87.18% (0.0000 - 0.0000)	87.11% (0.0000 - 0.0000)	86.67% (0.0000 - 0.0000)	86.01% (0.0000 - 0.0001)
10000	87.24% (0.0000 - 0.0000)	87.24% (0.0000 - 0.0000)	87.23% (0.0000 - 0.0000)	87.19% (0.0000 - 0.0000)	87.13% (0.0000 - 0.0000)

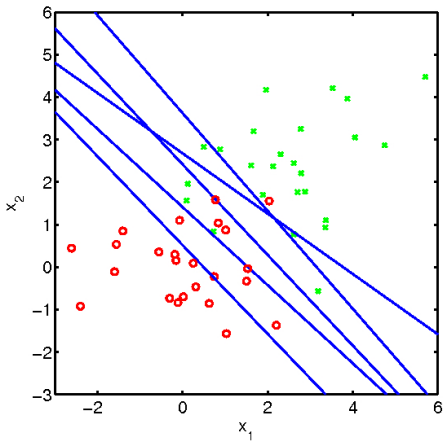
Table D.5: The relative difference in average variance between OC-SVM and SVM, $\frac{Var(SVM)-Var(OC-SVM)}{Var(SVM)}$ of 100 solutions. In parentheses the average variances of OC-SVM and SVM are given. The data are with noise of the size $\sigma = 0.6$.

$p\backslash n$	2	5	10	50	100
10	36.84% (0.0193 - 0.0309)	23.31% (0.0100 - 0.0130)	4.11% (0.0042 - 0.0044)	1.44% (0.0008 - 0.0008)	1.18% (0.0004 - 0.0004)
100	0.95% (0.0067 - 0.0067)	9.89% (0.0045 - 0.0050)	12.73% (0.0031 - 0.0035)	3.23% (0.0008 - 0.0008)	0.71% (0.0004 - 0.0004)
500	0.07% (0.0018 - 0.0018)	0.37% (0.0016 - 0.0016)	1.73% (0.0013 - 0.0014)	3.71% (0.0006 - 0.0006)	2.68% (0.0003 - 0.0003)
1000	0.01% (0.0010 - 0.0010)	0.06% (0.0009 - 0.0009)	0.38% (0.0008 - 0.0008)	7.48% (0.0005 - 0.0005)	2.27% (0.0003 - 0.0003)
10000	0.00% (0.0001 - 0.0001)	0.00% (0.0001 - 0.0001)	0.00% (0.0001 - 0.0001)	0.12% (0.0001 - 0.0001)	0.51% (0.0001 - 0.0001)

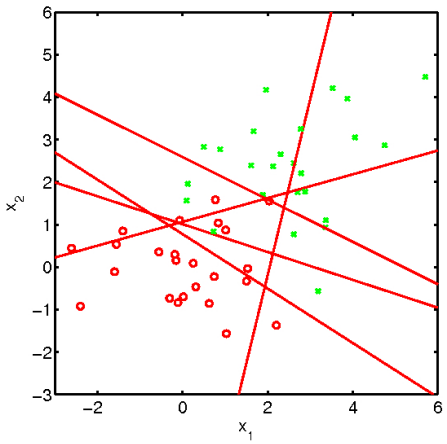


(a) OC-SVM, $\sigma = 0$

(b) SVM, $\sigma = 0$

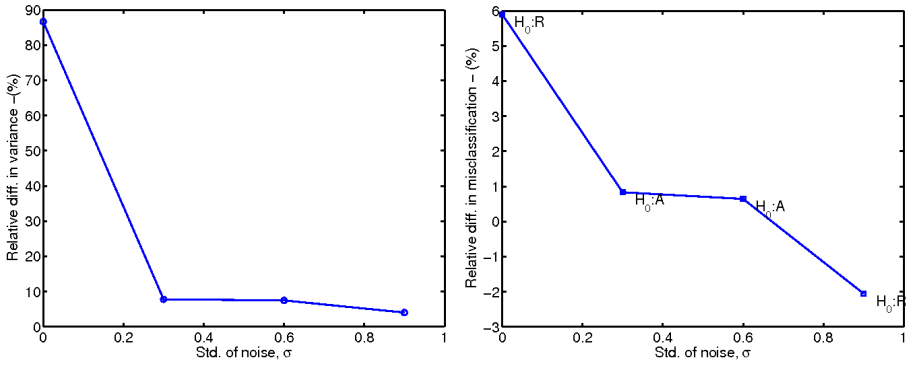


(c) OC-SVM, $\sigma = 0.9$



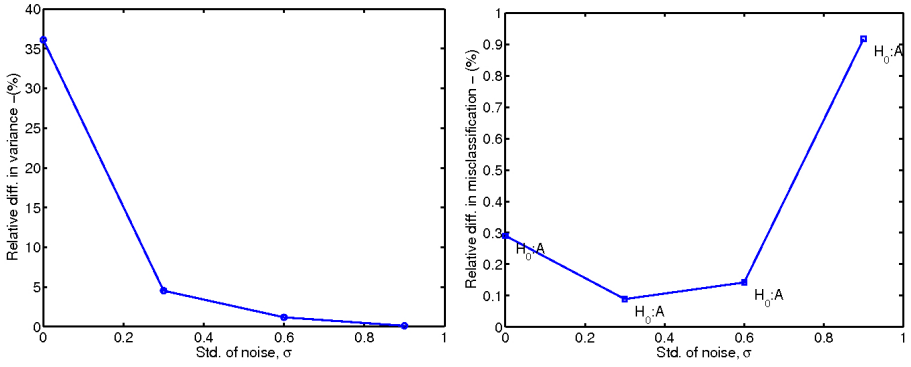
(d) SVM, $\sigma = 0.9$

Figure D.3: A visualization of the difference between OC-SVM and SVM on six different simulated data sets of 5 points in each group with no noise and noise with $\sigma = 0.9$ added to the translation, and a distance between the groups of 3 ($3\sqrt{\frac{1}{2}}$ in each direction).



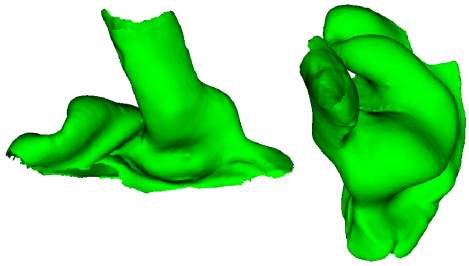
(a) Variance (b) Misclassification

Figure D.4: The average variance and misclassifications of the SVM and OC-SVM solutions as a function of the noise level for $p = 100$ and $n = 10$. (a) The relative difference between the variances is noted. (b) The result of test of the null-hypothesis that the means of the misclassification rates are equal is noted as R=rejected and A=accepted.

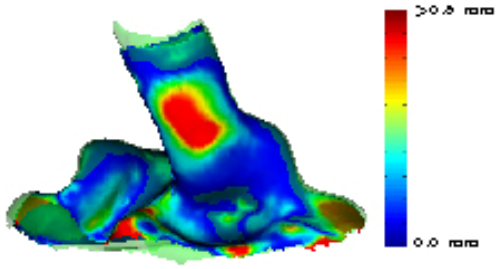


(a) Variance (b) Misclassification

Figure D.5: The average variance and misclassifications of the SVM and OC-SVM solutions as a function of the noise level for $p = 5$ and $n = 100$. (a) The relative difference between the variances is noted. (b) The result of test of the null-hypothesis that the means of the misclassification rates are equal is noted as R=rejected and A=accepted.



(a) Ear from 2 perspectives

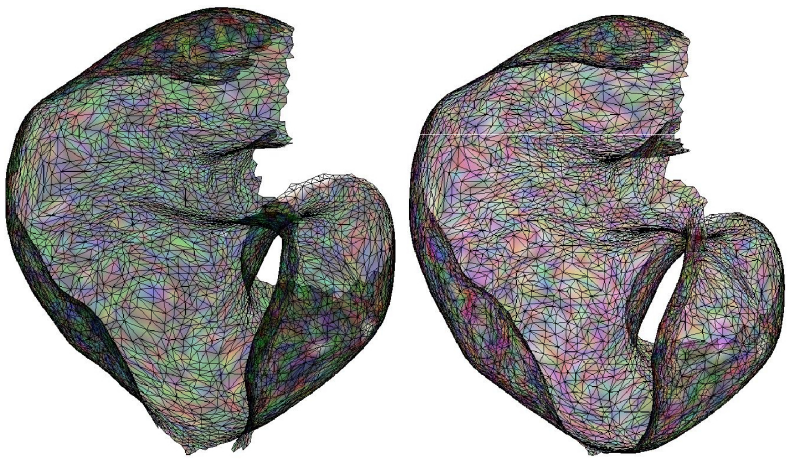


(b) Difference

Figure D.6: Typical example of an ear impression from two perspectives and an example of the difference map between an open and a closed mouth impression.

Table D.6: The relative difference in average variance between OC-SVM and SVM, $\frac{Var(SVM)-Var(OC-SVM)}{Var(SVM)}$ of 100 solutions. In parentheses the average variances of OC-SVM and SVM are given. The data are with noise of the size $\sigma = 0.9$.

$p \backslash n$	2	5	10	50	100
10	22.22% (0.0338 - 0.0435)	14.66% (0.0205 - 0.0240)	8.17% (0.0095 - 0.0103)	0.94% (0.0018 - 0.0018)	0.11% (0.0010 - 0.0010)
100	0.60% (0.0081 - 0.0081)	3.66% (0.0063 - 0.0066)	11.25% (0.0049 - 0.0055)	1.77% (0.0015 - 0.0016)	0.48% (0.0008 - 0.0008)
500	0.02% (0.0019 - 0.0019)	0.12% (0.0018 - 0.0018)	0.82% (0.0016 - 0.0017)	2.90% (0.0010 - 0.0010)	0.51% (0.0006 - 0.0006)
1000	-0.00% (0.0010 - 0.0010)	0.03% (0.0009 - 0.0009)	0.20% (0.0009 - 0.0009)	4.03% (0.0006 - 0.0007)	1.78% (0.0005 - 0.0005)
10000	-0.00% (0.0001 - 0.0001)	0.00% (0.0001 - 0.0001)	0.00% (0.0001 - 0.0001)	0.06% (0.0001 - 0.0001)	0.22% (0.0001 - 0.0001)



(a) SVM (b) OC-SVM

Figure D.7: The variance of the solutions, β projected onto the ear canal shape model. Results are shown for 500 SVM and OC-SVM solutions of the ear canal problem. The darker the color (RGB), the higher the variation of the parameters in $\boldsymbol{\beta}$. The variation of the SVM is twice that of the OC-SVM.

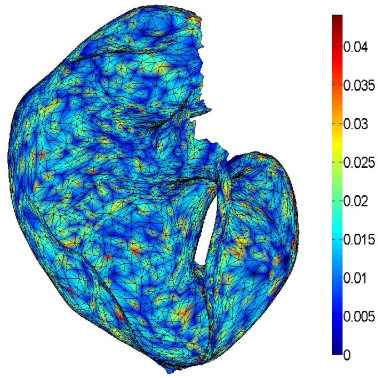


Figure D.8: A random normalized solution of the OC-SVM projected onto the mean shape of the ears, ie, the discriminant direction.

Appendix E

Multispectral recordings and analysis of psoriasis lesions

Authors: Line H. Clemmensen[1] and Bjarne K. Ersbøll[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

Published in proceedings *Workshop on Biophotonics Imaging for Diagnostics and Treatment, October 6, 2006 Proceedings, 9th MICCAI Conference*, 2006, p. 15-18.

E.1 Abstract

An objective method to evaluate the severeness of psoriasis lesions is proposed. In order to obtain objectivity multi-spectral imaging is used. The multi-spectral images give rise to a large p, small n problem which is solved by use of elastic net model selection. The method is promising for further studies of larger data sets including more patients than the four regarded here.

E.2 Introduction

Psoriasis is a disorder of excessive growth and reproduction of skin cells which may be caused by a immune-mediated disorder. The lesions caused by the excessive skin growth are red and often inflammatory. Furthermore, 1/3 of people with psoriasis report a family history of the disease. It affects both sexes and can occur at any age. The prevalence is estimated at 2-3% of Western populations.

The diagnosis of the severity of the psoriasis is important with regards to choice of treatment. A diagnosis cannot be performed by blood tests and the standard is to perform a visual diagnosis which requires trained staff. In some cases a biopsy is performed in order to rule out other disorders. An objective method for diagnosis would be less time consuming and less expensive. In addition to this, inter-observer variability can be avoided.

Traditionally, evaluation of psoriasis lesions are performed subjectively by trained staff using the PASI (*psoriasis area and severity index*, Fredriksson and Petersson (1978)). This evaluation form is limited with regards to large-scale studies. In 2001 SAPASI (*self-administrated* PASI) was proposed where the evaluation is performed by the patients themselves Szepietowski et al. (2001). This study concluded that objective methods for clinical evaluation of psoriasis is needed.

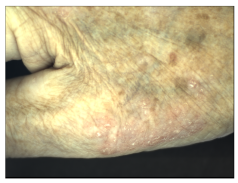
The ratings of the four patients considered here have been performed according to the severity index of the PASI. It's scale is from 0 (none) to 4 (maximum). The severity of the lesions are measured by the degree of erythema and the degree of infiltration of the lesions. Erythema is the redness of the skin caused by dilatation and congestion of the capillaries. This is often a sign of inflammation or infection. Infiltration refers to the thickness of the psoriasis lesion[1].

To obtain an objective method of evaluation multi-spectral imaging is considered. Each multi-spectral image consists of nine spectral bands. Hence, a large amount of data is present for each of the few observations. Such constitutions are referred to as large p, small n problems. To analyze the problem at hand we use *least angle regression - elastic net* which introduces a sparsity into the solution and in this way selects a subset of features Zou and Hastie (2005).

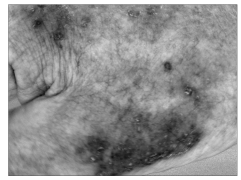
[1] Psoriatic skin is thicker than healthy skin, Fredriksson and Petersson (1978).

E.3 Method

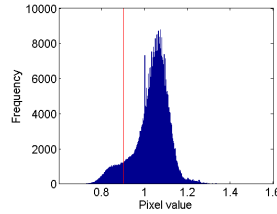
This study considers four patients each with two lesions imaged. Two to five images have been acquired of each lesion area. This amounts to a total of 26 images. The lesions have been valuated in the range from 0 to 2, i.e. the variance regarding the severity index is small within the four patients. The segmentation of the ROIs (*regions of interest*) of the inflammations and the scales of the lesions is illustrated in Figure E.1.



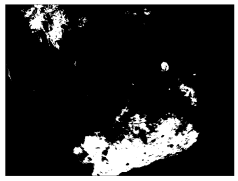
(a) RGB representation.



(b) Relation between 5th and 7th spectral bands.

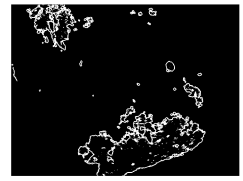


(c) Histogram of (b) with threshold 0.9 for inflammation ROI.



(d) ROI for inflammation obtained from (c).

(e) Dilation of (d) with a disk kernel of size 8 and erosion with 5.

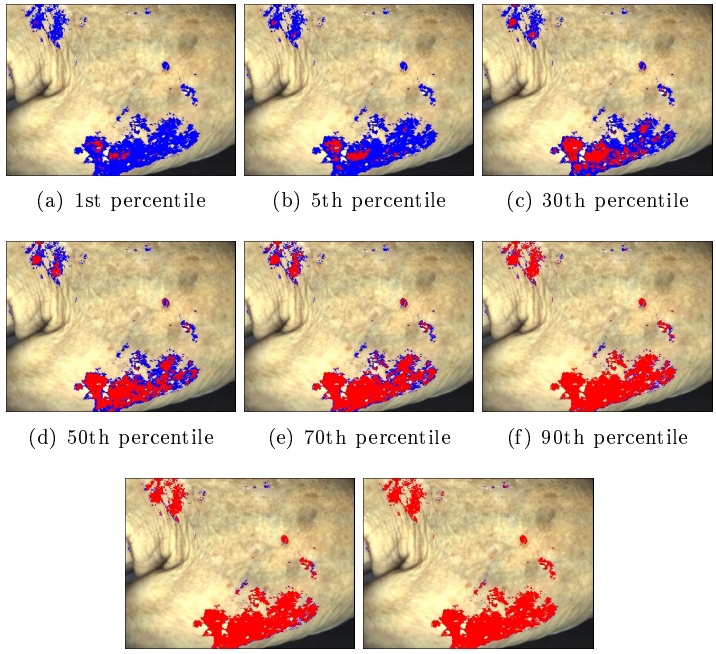


(e)-(d).

(f) ROI for scales obtained as

Figure E.1: Illustration of the segmentation of the ROIs in the images. The pairwise relation between the 5th (amber) and 7th (red) spectral bands, (b) is used since this emphasizes the red inflammations. Two ROIs are segmented: One containing the inflammation (d) and another containing the scales (f).

From the original spectral bands and from the pairwise ratios between the spectral bands the following features are extracted from both the inflammation ROI and the scale ROI: The 1st, 5th, 30th, 50th, 70th, 90th, 95th, and 99th percentiles. This amounts to 1458 features. Figure E.2 illustrates the spatial distribution of the pixels in relation to the percentiles of the pairwise differences between the amber (592nm) and the res (630nm) bands. It is seen that per-



(a) 1st percentile (b) 5th percentile (c) 30th percentile

(d) 50th percentile (e) 70th percentile (f) 90th percentile

(g) 95th percentile (h) 99th percentile

Figure E.2: Illustration of the percentile features for one of the spectral images. The region of interest is marked with blue and the red pixels are the pixels which have a value lower than the given percentile. The region and the pixels are illustrated on top of a pseudo RGB image of the corresponding sample.

centiles contain a certain information of the spatial distribution of the pixels, and thus not all spatial information is lost when summary statistics are used to sum up the information in the region of interest. Furthermore, the summary statistics in general serve as more robust measures of the pixel values than the pixel values in themselves, i.e. some of the natural biological variation between pixel values are averaged out by looking at the distribution of the pixel values. This exploits the first of the blessings in high dimensional problems, namely that we can average over similar features and thereby obtain fewer dimensions and more robust estimates/features.

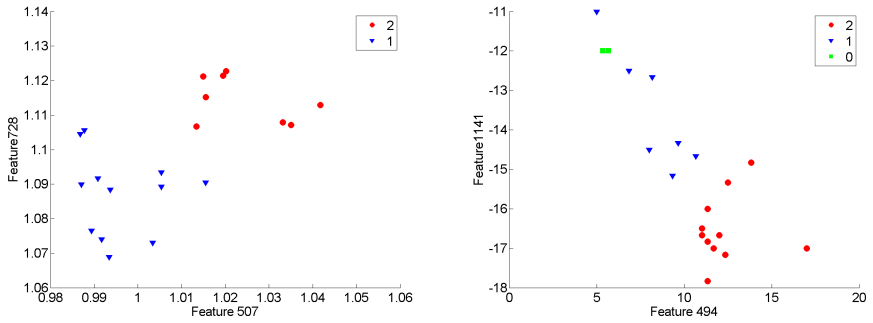
LARS-EN (*least angle regression - elastic net*) model selection, proposed in Zou and Hastie (2005), combines Ridge regression Hoerl and Kennard (1970) and Lasso model selection Tibshirani (1996); Efron et al. (2004) and hereby obtains sparse solutions with the computational effort of a single ordinary least squares fit. This method is used to analyze the large p, small n problem at hand.

E.4 Results and discussion

Patient number three is not included in the analysis since both the RGB image and the further analysis imply that this is an outlier. The analysis is performed using LARS-EN with leave-one-out cross-validation Hastie et al. (2009). Only two features are needed to describe the degree of erythema and of infiltration, respectively. The features are listed in table E.1. The results are illustrated in Figure E.3. The variables give a good ordering of the evaluations. Furthermore, the standard deviations for the training and the test are: 0.4/0.5 and 0.5/0.6 for erythema/infiltration.

Table E.1:

Feature	Description (Eurythema)
507	10th percentile of the inflammation ROI in the ratio of the blue (472nm) and green (515nm) bands
728	95th percentile of the inflammation ROI in the ratio of the cyan (503nm) and green (515nm) bands
Feature	**Description (Infiltration)**
494	95th percentile of the inflammation ROI in the difference of the blue (472nm) and green (515nm) bands
1141	90th percentile of the inflammation ROI in the difference between the amber (592nm) and red (630nm) bands



(a) Erythema, evaluations: {1,2}. (b) Infiltration, evaluations: {0,1,2}.

Figure E.3: Scatter plots of the two most frequently selected variables with leave-one-out cross-validation for erythema and infiltration, respectively.

Summing up, the results are promising as variables are selected which give a good ordering of the patients according to the severity index ratings. Furthermore, the standard deviations of the leave-one-out cross-validation are relatively small for only two variables. The next step will be to evaluate the method on larger data sets.

E.5 Acknowledgements

The authors would like to thank dermatologist Dr. Lone Skov at Gentofte Hospital, Denmark for her cooperation and for performing the severity index evaluations of the four patients.

Appendix F

Individual discriminative face recognition models based on subsets of features

Authors: Line H. Clemmensen[1] and David D. Gomez[2] and Bjarne K. Ersbøll[1].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

2. Computational Imaging Lab, Pompeu Fabre University, Barcelona, Spain.

Published in *SCIA 2007 LNCS 4522 proceedings*, p. 61-71, Ed. 2, Springer-Verlag, 2007.

F.1 Abstract

The accuracy of data classification methods depends considerably on the data representation and on the selected features. In this work, the elastic net model selection is used to identify meaningful and important features in face recognition. Modelling the characteristics which distinguish one person from another using only subsets of features will both decrease the computational cost and increase the generalization capacity of the face recognition algorithm. Moreover, identifying which are the features that better discriminate between persons will also provide a deeper understanding of the face recognition problem. The elastic net model is able to select a subset of features with low computational effort compared to other state-of-the-art feature selection methods. Furthermore, the fact that the number of features usually is larger than the number of images in the data base makes feature selection techniques such as forward selection or lasso regression become inadequate. In the experimental section, the performance of the elastic net model is compared with geometrical and color based algorithms widely used in face recognition such as Procrustes nearest neighbor, Eigenfaces, or Fisherfaces. Results show that the elastic net is capable of selecting a set of discriminative features and hereby obtain high classification rates.

F.2 Introduction

Historical facts (New York, Madrid, London) have put a great emphasis on the development of reliable and ethically acceptable security systems for person identification and verification. Traditional approaches such as identity cards, PIN codes, and passwords are vulnerable to falsifications and hacking, and such security breaks thus also appear frequently in the media.

Another traditional approach is biometrics. Biometrics base the recognition of individuals on the intrinsic aspects of a human being. Examples are fingerprint and iris recognition Daugman (2002)Daugman (1993). However, traditional biometric methods are intrusive, i.e. one has to interact with the individual who is to be identified or authenticated. In some cases, however, iris recognition is implemented as a standard security check in airports (e.g. New York JFK). Recognition of people from facial images on the other hand is non-intrusive. For

this reason, face recognition has received increased interest from the scientific community in the recent years.

Face recognition consists of problems with a large number of features (of geometrical or color related information) in relation to the number of face images in the training sets. In order to reduce the dimensionality of the feature space we propose to use *least angle regression - elastic net* (LARS-EN) model selection to select discriminative features that increase the accuracy rates in facial identification. LARS-EN was introduced by Zou et. al in 2005 Zou and Hastie (2005). It regularizes the *ordinary least squares* (OLS) solution with both the Ridge regression and Lasso constraints. The method selects variables into the model where each iteration corresponds to loosening the regularization with the Lasso constraint. The ridge constraint ensures that the solution does not saturate if there are more variables in the model than the number of observations.

The rest of the paper is organized as follows: In section two, a review of the standard face recognition techniques is presented. Section three describes the LARS-EN algorithm. In section four, we describe and state the results for several experiments which we conducted to test the discriminative capacity of the obtained features. Finally, section 5 gives a conclusion of the conducted experiments and discusses some future aspects of the research.

F.3 Face recognition review

The first techniques developed for face recognition aimed at identifying people from facial images based on geometrical information. Relative distances between key points such as mouth or eye corners were used to characterize faces Goldstein et al. (1971)Craw et al. (1999). At this first stage of facial recognition, many of the developed techniques focused on automatic detection of individual facial features. The research was notably strengthened with the incursion of the theory of statistical shape analysis. Within this approach, faces were described by landmarks or points of correspondence on an object that matches between and within populations. In a 2D-image, a landmark $\mathbf{l}$ is a two dimensional vector $\mathbf{l} = (x, y)$ that, to obtain a more simple and tractable mathematical description, is expressed in complex notation by $\mathbf{l} = x + iy$, where $i = \sqrt{-1}$. In this framework, a face in an image is represented by a configuration or a set of n landmarks $[\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n]$ placed on meaningful points. Geometrical face recognition based on landmarks is conducted by evaluating the similarity of the configuration of a test face with respect to the configurations in a facial database. In order to

achieve this, different measures of similarity have been proposed, see e.g. Dryden and Mardia (1998). Among all the proposed metrics, the Procrustes distance has been the most frequently used. Given two configurations w and z, the Procrustes distance between them is defined by

$$D_P(w,z) = \inf_{\beta,\theta,a,b} \| \frac{z}{\|z\|} - \frac{w}{\|w\|}\beta e^{i\theta} - a - ib \| \quad , \tag{F.1}$$

where $\| \cdot \|$ represents the l_2 norm, and the parameters β, θ, a, and b, which denotes a scaling, a rotation, and a translation of configuration w, are chosen to minimize the distance between w and z. Several extensions of this measure have been proposed. For instance, Shi et. al Shi et al. (2006) has recently proposed a refined Procrustes distance based on principal component analysis. The configurations (the landmark representations of the faces) are first centered at the origin and transformed to have unit size. Then a complex principal component analysis is conducted to reduce the dimensionality. The similarity measure is defined in this lower m-dimensional space by

$$D_{RP}(w,z) = \sum_{k=1}^{m} \| \frac{\hat{z}_k}{\sqrt{\lambda_k^{(z)}}} - \frac{\hat{w}_k}{\sqrt{\lambda_k^{(w)}}} \| \quad , \tag{F.2}$$

where $\hat{z}_k$ is the k^{th} eigenvector of configuration y, $\hat{w}_k$ is the k^{th} eigenvector of configuration w, and $\lambda_k^{(z)}$ and $\lambda_k^{(w)}$ the corresponding eigenvalues.

The publication of Eigenfaces by Turk and Pentland Turk and Pentland (1991) showed that it was possible to obtain better classification rates by using the color intensities. Since then, geometrical face recognition was gradually declining until the extent that, nowadays, it principally remains to support color face recognition. The appearance of Eigenfaces provided an excellent way of summarizing the color information of the face. The facial images in a training database were first registered to obtain a correspondence of the pixels between the images. Then, a principal component analysis was conducted to reduce the high data dimensionality, to eliminate noise, and to obtain a more compact representation of the face images. When a new test image was desired classified, the same data reduction was applied to obtain a comparable compact test image representation. The similarity of the compact test image representation was measured with each of the compact training image representations based on the Euclidean distance. The test image was associated with the training image with the smallest Euclidean distance. Based on Eigenfaces, Fisherfaces obtained higher classification rates by applying a Fisher Linear discriminant on

the obtained principal components. As a result of the publication of Fisherfaces a considerable percentage of the current research in the field is devoted to find more discriminative projections Belhumeur et al. (1997)Cevikalp et al. (2005).

In this paper, an approach to increase the discrimination among individuals is proposed. However, instead of looking for more discriminative projections as the previous methods, it aims at finding more discriminative features. This is in line with the face detector of Viola and Jones Viola and Jones (2001) that selects Haar features which are important for the face detection task. Basing the identification on only a subset of the features will make the system work faster for future identifications. The approach is described in next section.

F.4 Elastic net model selection

We consider the linear model:

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon} \quad , \tag{F.3}$$

where each $\epsilon_i \sim N(0, \sigma^2)$. We assume $\vec{y}$ centered (i.e. $\sum_{i=1}^{n} y_i = 0$) and the columns of $\mathbf{X}$ normalized to zero mean and unit length.

The LARS-EN method is used to make multiple individual discriminative models by the use of dependent variables with ones and zeros discriminating one individual from the remaining people in the data set. In the case of one image per individual the k^{th} individual model is:

$$\text{center}\left(\begin{bmatrix} \vec{0}_{k-1} \\ 1 \\ \vec{0}_{n-k} \end{bmatrix}\right) = \text{normalize}\left(\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}\right)\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad , \tag{F.4}$$

where n is the number of individuals (there are $n-1$ individuals distinct from individual k), and p is the number of features. $\vec{0}_{k-1}$ denotes a vector of $k-1$ zeros. The geometrical features used in this work were the x and the y coordinates of the landmarks. The color based features were the gray scale intensities of the facial images after warping.

F.4.1 The elastic net

Least angle regression - elastic net (LARS-EN) model selection was proposed by Zou et. al Zou and Hastie (2005) to handle $p \gg n$ problems. The method

regularizes the *ordinary least squares* (OLS) solution using two constraints, the 1-norm and the 2-norm of the coefficients. These constraints are the ones used in the *least absolute shrinkage and selection operator* (Lasso) Tibshirani (1996) and Ridge regression Hoerl and Kennard (1970), respectively. The naive elastic net estimator is defined as

$$\hat{\vec{\beta}} = \text{argmin}_\beta \{ \|\vec{y} - \mathbf{X}\vec{\beta}\|_2^2 + \lambda_1 \|\vec{\beta}\|_1 + \lambda_2 \|\vec{\beta}\|_2^2 \} \quad , \tag{F.5}$$

where $\|\vec{\beta}\|_1 = \sum_{i=1}^p |\beta_i|$, $|\cdot|$ denoting the absolute value, and $\|\vec{\beta}\|_2^2 = \sum_{i=1}^p \beta_i^2$. Choosing $\lambda_1 = 0$ yields Ridge solutions, and likewise choosing $\lambda_2 = 0$ yields Lasso solutions. For the Lasso method it is likely that one or more of the coefficients is zero at the solution, while for the Ridge regression it is not very likely that one of the coefficients is zero. Hence, we obtain a sparsity in the solution by using the Lasso constraint. The Ridge constraint ensures that we can enter more than n variables into the solution before it saturates.

We can transform the naive elastic net problem into an equivalent Lasso problem on the augmented data (c.f. (Zou and Hastie, 2005, Lemma 1))

$$\mathbf{X}^* = (1+\lambda_2)^{-1/2} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix} \quad , \quad \vec{y}^* = \begin{bmatrix} \vec{y} \\ \vec{0}_p \end{bmatrix} \quad . \tag{F.6}$$

The normal equations, yielding the OLS solution, to this augmented problem are

$$\begin{aligned} \left(\frac{1}{\sqrt{1+\lambda_2}}\right)^2 \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix} \hat{\vec{\beta}}^* &= \frac{1}{\sqrt{1+\lambda_2}} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix}^T \begin{bmatrix} \vec{y} \\ \vec{0}_p \end{bmatrix} \Leftrightarrow \\ \frac{1}{\sqrt{1+\lambda_2}} \left(\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I}_p^T \mathbf{I}_p\right) \hat{\vec{\beta}}^* &= \mathbf{X}^T \vec{y} \quad . \end{aligned} \tag{F.7}$$

We see that $\frac{1}{\sqrt{1+\lambda_2}}\hat{\vec{\beta}}^*$ is the Ridge regression estimate with parameter λ_2. Hence, performing Lasso on this augmented problem yields an elastic net solution. The *least angle regression* (LARS) model selection method proposed by Efron et al. (2004) can be used with advantage to compute the Lasso solution on the augmented problem. The LARS algorithm obtains the Lasso solution with a computational speed comparable to computing the OLS solution of the full set of covariates.

The algorithm uses the LARS implementation with the Lasso modification as described in the following section. Hence, we have the parameter λ_2 to adjust, but also the number of iterations for the LARS algorithm can be used. The larger λ_2, the more weight is put on the Ridge constraint. The Lasso constraint is weighted by the number of iterations. Few iterations corresponds to a high value of λ_1, and vice versa. The number of iterations can also be used to ensure a low number of active variables like the forward selection procedure.

F.4.2 Least angle regression

The least angle regression selection (LARS) algorithm method proposed by Efron et. al Efron et al. (2004) finds the predictor most correlated with the response, takes a step in this direction until the correlation is equal to another predictor, then it takes the equiangular direction between the predictors of equal correlation (*the least angle direction*) and so forth.

By ensuring that the sign of any non-zero coordinate β_j has the same sign as the current correlation $\hat{c}_j = \vec{x}_j^T(\vec{y} - \hat{\vec{\mu}})$, the LARS method yields all Lasso solutions[1]. This result is obtained by differentiating the Lagrange version of the Lasso problem. For further details see Efron et al. (2004).

F.4.3 Distance measure

By introducing a distance measure we obtain a measure of how close a new image is to the different individuals in the database. We used the absolute difference between the predicted value $\hat{y}_k$ for model k and the true value y_k for an image belonging to individual k as a measure of the distance between the new image and individual k.

F.5 Results and comparison

In order to test the performance of LARS-EN with respect to the previously commented geometrical and color face recognition technique, two identification experiments were conducted. The difference of the experiments is in the used features. In the first experiment, only the landmarks were used. The second experiment considered only the color. In order to conduct the experiments, the XM2VTS database was used Messer et al. (1999). Eight images for each of the first 50 persons were selected. For all experiments a 4-2-2 strategy was chosen: 4 images of each person to train the model, 2 images of each person to adjust the parameters in the model, and 2 images of each person to verify the model.

To evaluate the performance of the algorithms we used rank plots of the cumulative match scores as proposed in Philip et al. (2000). The horizontal axis of the rank plots is the rank itself (referring to the sorted distance measure) and

[1] $\vec{y}$ is centered and normalized to unit length, $\mathbf{X}$ is normalized so each variable has unit length, and $\hat{\vec{\mu}} = \mathbf{X}\hat{\vec{\beta}}$.

the vertical axis is the cumulated probability of identification. Hence, we obtain an answer to the question: "Is the correct match in the top n matches?".

F.5.1 Geometrical face recognition

In order to conduct this first experiment, a set of 64 landmarks were placed along the face, eyes, nose and mouth of each of the 400 selected images. Figure F.1 displays the landmarks used in the experiment.

Figure F.1: Illustration of the landmarks used in the experiment.

Table F.1 summarizes the classification rates obtained using only the landmarks. The LARS-EN method has higher classification rates than Procrustes, Refined Procrustes, and PCA, but not the Fisher method.

Method/Classification rate	Training	Validation	Test
Procrustes	1.00	-	0.67
Refined Procrustes	1.00	0.76	0.52
PCA	1.00	0.73	0.63
PCA+Fisher	1.00	0.88	0.81
LARS-EN	0.96	0.76	0.71

Table F.1: Summary of the classification rates for the models based solely on the landmarks.

The LARS-EN models included on average 52 of the 128 shape features (x and y coordinates of the landmarks). It should be noted that the mean square error of both the training and the test set in LARS-EN were of the same size,

i.e. no severe overfitting was observed. Furthermore, LARS-EN seems to be more honest in the training error and in that sense overfit less than the other methods compared. Figure F.2 illustrates a rank plot of the performances of the landmark models. We see a good performance for LARS-EN better than PCA, Refined Procrustes, and Procrustes, and also based on fewer features.

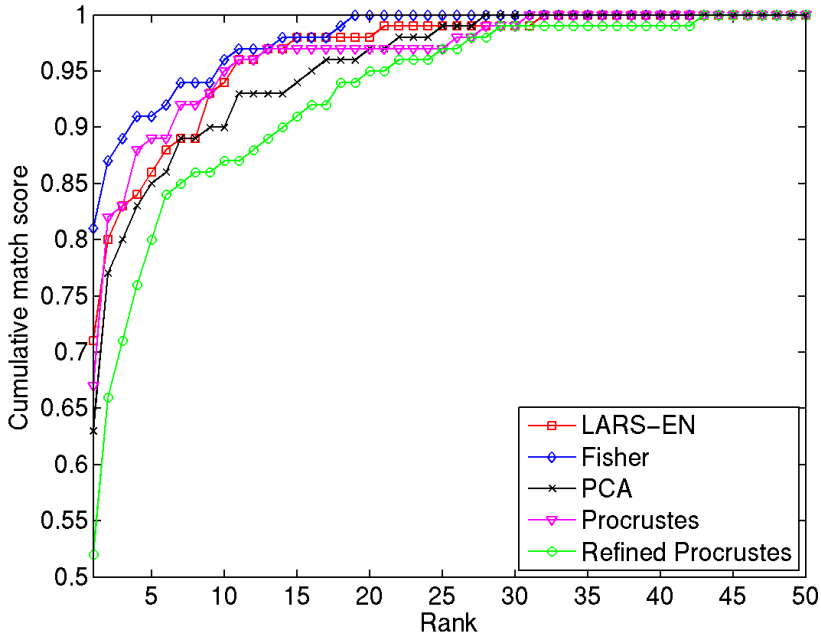


Figure F.2: Identification performance of the models based solely on landmarks.

Figure F.3 illustrates which landmarks are selected for four of the individual models. Observe how the selected landmarks depend on the facial characteristics of each person.

F.5.2 Color face recognition

In order to obtain a one to one correspondence of pixels between the images the faces were aligned with warping. The same 4-2-2 validation strategy as before was applied and the Eigenfaces, Fisherfaces, and LARS-EN methods were compared. Table F.2 summarizes the results.

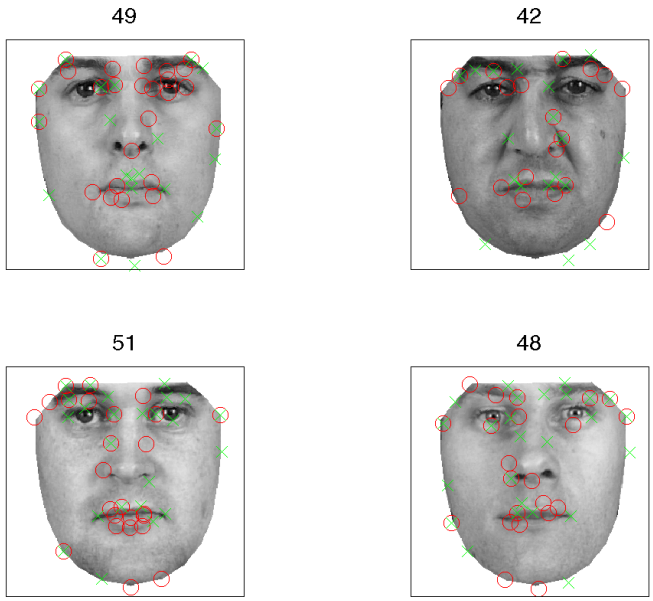


Figure F.3: Illustration of four persons and the selected landmarks in the individual LARS-EN models. x-coordinates are marked with crosses, and y-coordinates are marked with circles. From left to right the person are: No. 1, no. 13, no. 36, and no. 44.

Method/Classification rate	Training	Validation	Test
Eigenfaces	1	0.87	0.85
Fisherfaces	1	0.96	0.94
LARS-EN	1	0.97	0.92

Table F.2: Summary of the classification rates for the models based solely on the color information.

Based on color information we observed higher classification rates than those for LARS-EN based on geometrical information. LARS-EN and Fisherfaces were comparable while both were better than Eigenfaces. The LARS-EN models included around 2000 features (pixels) out of approximately 47000. Figure F.4 illustrates the performance of the color based methods. The performance

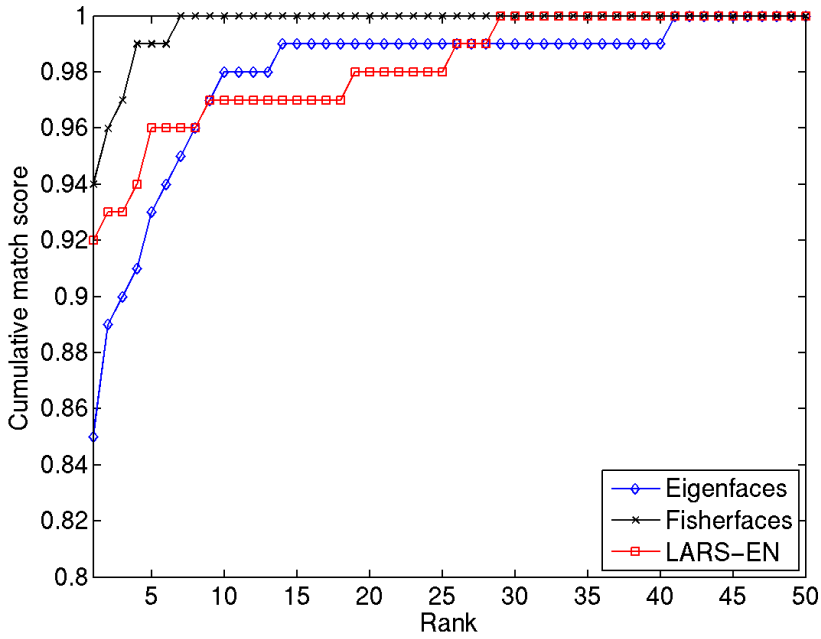


Figure F.4: Identification performance of the models based solely on color information.

of Fisherfaces was slightly better than for the other two methods which were comparable in performance.

Similar to what was done for the geometric features we now examine which features were selected in experiment two. Figure F.5 shows the selected color pixels on four different persons. The selected pixels are to a high degree situated around the eyebrows, the eyes, the nose, and the mouth, but also on e.g. the cheeks and the chin. Furthermore, the features are individual from person to person. Observe e.g. the different selection of pixel features on and around the noses of the individuals.

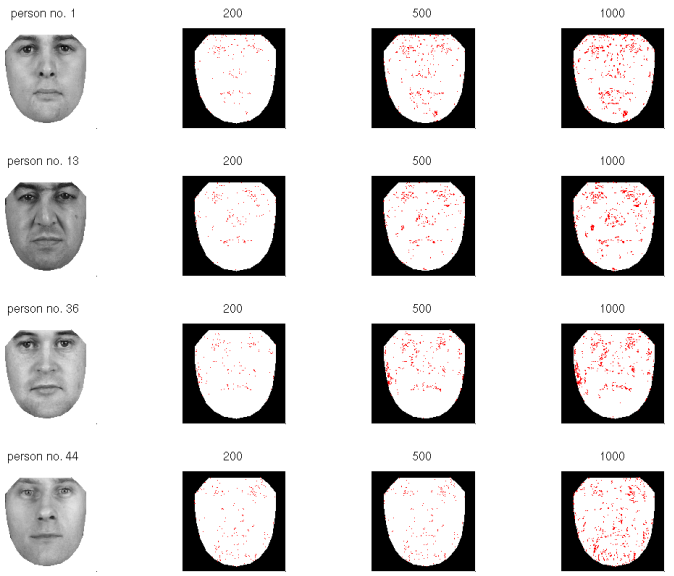


Figure F.5: Illustration of four persons with the first 200, 500, and 1000 selected pixels marked.

F.6 Discussion and conclusion

The LARS-EN method performed better than the reference methods Procrustes, refined Procrustes, and PCA, but nor better than PCA+Fisher when based solely on information from landmarks.

Based on color information the LARS-EN models obtained better classification rates than the Eigenfaces and classification rates comparable to Fisherfaces.

Additionally, we identified important features via the feature selection. For the landmarks, only 52 features were needed on average for the individual models. The color models were based on around 2000 features which were situated around the eyes, the nose, the mouth, and the eyebrows, but also on the cheeks and the chin. The selected features differ from individual to individual. Furthermore, the reduction of the feature space decreases the computational efforts for predictions.

Consequently, our results show that a limited number of geometrical or color features can suffice for face recognition, and emphasize that geometrical information should not be disregarded. There are several other possibilities of feature extraction from geometrical information of faces, such as ratios and angles between landmarks, which would be interesting to explore. The LARS-EN algorithm is a good tool for exploring new feature spaces and finding the more interesting ones.

In future work, it is furthermore of interest to examine the methods for a larger database.

F.7 Acknowledgements

The authors would like to thank Karl Sjöstrand who has implemented the LARS and LARS-EN methods in Matlab. The implementations are available at his homepage[2].

[2] www.imm.dtu.dk/~kas

Appendix G

Temporal reflectance changes in vegetables

Authors: Bjørn S. Dissing[1] and Line H. Clemmensen[1] and Hanne Løje[2] and Bjarne K. Ersbøll[1] and Jens Adler-Nissen[2].

1. Informatics and Mathematical Modelling, Technical University of Denmark.

2. National Food Institute of Denmark, Technical University of Denmark.

Published in *CRICV 2009 proceedings*, IEEE, 2009.

G.1 Abstract

Quality control in the food industry is often performed by measuring various chemical compounds of the food involved. We propose an imaging concept for acquiring high quality multispectral images to evaluate changes of carrots and celeriac over a period of 14 days. Properties originating in the surface chemistry of vegetables may be captured in an integrating sphere illumination which enables the creation of detailed surface chemistry maps with a good combination of spectral and spatial resolutions. Prior to multispectral image recording, the vegetables were pre-fried and frozen at $-30°C$ for four months. During the 14 days of image recording, the vegetables were kept at $+5°$ C in refrigeration. In this period, surface changes and thereby reflectance properties were very subtle. To describe this small variation we employed advanced statistical techniques to search a large featurespace of variables extracted from the chemistry maps. The resulting components showed a change in both the carrot and celeriac samples. We were able to deduct from the resulting components that oxidation caused the changes over time.

G.2 Introduction

Quality assessment of food products is a non trivial task which has been approached in different ways over time. Depending on the food product, different parameters are considered important for the overall quality estimation of the food product. Parameters such as surface color, texture and appearance are very general, and should be assessed in most quality estimation scenarios.
Online quality inspection for food process control is today often done by human expert operators who have many years of experience. However, the trend seems to point towards fast non-invasive inspection methods such as Near Infra Red (NIR) technology for quality inspection in different food process control tasks instead. We propose the use of multispectral imaging in the visible as well as the NIR area of the electromagnetic spectrum to quantify chemical properties of food, and thereby stating its level of quality, instead of human operators and as an alternative to standard NIR measurement methods. By employing imaging instead of point measurements it is possible to gain more spatial information about the process, which makes it possible to assess non-chemical as well as chemical quality features. Non-chemical quality features are evaluations of e.g.

piece-size, shape and texture.
In this study, we are specifically investigating the quality loss of meal elements for professionally prepared meals with regards to change in surface color after super-chilling and during thawing at +5°C over a period of 14 days. Meal elements are robust semi-prepared convenience components based typically on meat, fish or vegetables and meant for professional use. The authors have recently shown that pre-fried vegetable meal elements have promising properties with respect to high culinary quality and robustness towards freezing and thawing, thereby potentially solving a major hindrance for the use of heat treated vegetables as meal elements Adler-Nissen (2007). Super chilling involves a partial freezing of the products, which slows down quality deterioration Bao et al. (2007). In Vina and Chaves (2003) an experiment of celeriac stored in an refrigerated environment was carried out and various parameters were measured using traditional methods. Celeriac and carrots were the subjects of this study, where we measured the reflectance properties using a multispectral imaging device called VideometerLab which will be described in the next section. The pre-fried vegetables were produced by a new process for continuous stir-frying in industrial scale, which has been introduced for producing convenience high-quality vegetables Adler-Nissen (2002). The pre-fried vegetables have a low fat content (typically 1%-2% of the product weight), a texture and flavor similar to what can be achieved in the kitchen, and vitamins are preserved almost 100% Burgaard et al. (2004). In subsequent studies it was observed that the products may be frozen and re-heated on a frying pan or in a convection oven without any exudation of excess water, which is a major advantage over existing quick-frozen vegetables Adler-Nissen (2005).

G.3 Materials and methods

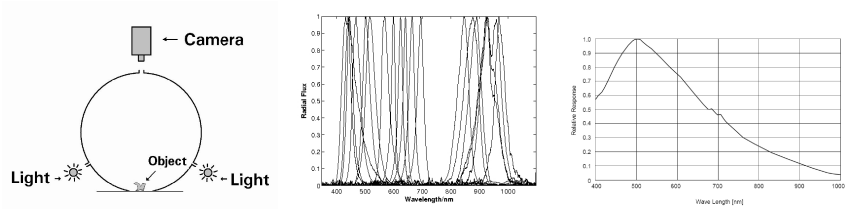
In the following, the experimental design, the acquisition of digital images of the vegetables and further post-processing of these are described.

G.3.1 Experimental setup

In the present work the quality of pre-fried vegetables (celeriac and carrots shaped as cubes of size approximately 0.5 cm^3) were evaluated (e.g. by means of change in color surface) after freezing and thawing. In a pilot plant, the raw products were pre-fried using a special frying machine "the continuous wok" Adler-Nissen (2002). After frying, the products were packed in 500g portions in plastic bags and frozen at $-30°C$. After four months of freezing, the bags with

the pre-fried vegetables were removed from the freezer and thawed up to 14 days at $+5°C$ in refrigeration. On each day of analysis (day 2, 4, 8, 10, 12 and 14) one plastic bag was taken out from the refrigerator and the contained vegetables were digitized. For both types of vegetables, the samples were digitized using two petri dishes to create a test and training set. The multispectral images were segmented in two steps, first isolating all vegetable-piece in the image, and then separating the pieces from each other. After segmentation, ratios were calculated for all combinations of wavelengths to remove shadow effects and possibly get better baseline separation in different spectral bands. For each ratio in each vegetable-piece, the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles were calculated. This yielded a total of 3249 variables, in a test and training set having 193 and 192 observations respectively for the carrot data. For the celeriac data, similar datasets were created yielding a total of 3249 variables with 207 and 206 observations in the test and training set respectively. Obviously we need a way to figure out which ratios best describe changes over time. For this task a penalized LS algorithm called LARS-EN which is described later was employed to find a set of optimal components. Subsequently statistical tests were performed to evaluate if the identified changes were significant.

G.3.2 VideometerLab



(a) (b) (c)

Figure G.1: a) Principle of imaging with integrating (Ulbricht) sphere illumination. The LEDs located in the rim of the sphere ensures narrowband illumination. b) Normalized spectral power distributions of the LEDs located in the VideometerLab. c) Spectral sensitivity of the camera mounted in VideometerLab. It is seen in a) that the camera is placed above the object of interest.

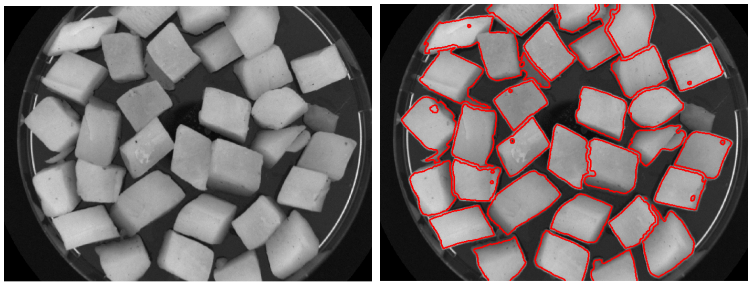
The acquisition of data was done using VideometerLab(http://www.videometer.com) which acquires multi-spectral images in up to 20 different wavelengths ranging from 430 to 970 nm. The camera setup is seen in figure G.1a. The object, in this paper, vegetables, is placed inside an integrating or Ulbricht sphere which has its interior coated to obtain high diffuse reflectivity for optimal light conditions. In the top of the sphere a camera is located with the sensitivity spectrum seen

in Figure G.1c. The sensitivity decays towards the near-infrared area, which means that the illuminating diodes in this area needs more power to achieve the same level of intensity as the visible bands. The LEDs, having the spectral radiant power distributions seen in figure G.1b, are strobing successively, resulting in an image for each LED of dimensionality 1280x960. These are calibrated radiometrically as well as geometrically to obtain the optimal dynamic range for each LED as well as to minimize distortions in the lens and thereby pixel-correspondence across the spectral bands. The well defined and diffuse illumination of the optically closed scene aims to avoid shadows and specular reflections. Furthermore, the system has been developed to guarantee the reproducibility of the collected images. This allows for comparative studies of time series of images Gomez et al. (2007).

G.3.3 Segmentation of the images

In the experiment we considered one vegetable-piece as an observation. In order to extract each vegetable-piece of the multispectral image seen in figure G.2, a relative difficult segmentation problem is at hand. This is caused by the fact that the individual pieces were not placed in a systematic manner where they were isolated, but instead lie in lumps, touching each other. This means that they cast shadows on each other as seen in figure G.2. Furthermore the pieces have very similar spectral fingerprints which means they cannot be discriminated using purely spectral values.
As an initial step, the background, meaning everything but the vegetables was isolated. This was done using Otsu's method Otsu (1979) on the multispectral image, projected onto a hyperplane. The projection function used to carry out this projection, optimally separates pixels coming from one of two populations. These populations are described by two types of labels which were manually annotated. One set of labels contained spectra of petri dish and general background, while the other of vegetable surface, either carrot or celeriac. The labeled data was used to calculate the projection function by means of a Canonical Discriminant Analysis (CDA) Hastie et al. (2009).
In order to isolate each piece, spatial information is needed, especially gradient information. This information acts as a good guide for the segmentation, and together with morphological transformations employed under a marker-controlled Watershed segmentation as described in Gonzalez et al. (2002) we were able to do an automatic segmentation of the vegetables. The result of the segmentation is seen in figure G.2. This seems to be a relatively good result, although flaws are present. In the upper left corner of the petri dish, two pieces are merged together as a result of bad gradient information. In the middle right side it is seen how a dim piece has been totally ignored by Otsu's method due to its dark

appearance. These flaws might be avoided by using an alternative segmentation technique but they were not crucial for the task at hand.



(a) (b)

Figure G.2: Both figures show band 10, corresponding to 645 nm of the carrot sample, on day 2. a) is the pure image and b) has the segmentation result superimposed on the image.

G.3.4 Feature selection method

Having a feature space with n observations and p variables, there are different ways of using this space to describe a variable depending on it. A common technique used to relate the dependent variable and the feature space in a well-posed problem is by using Ordinary Least Square (OLS). Here we have chosen the dependent variable to be the number of the day the observation belongs to, while the independent variables as mentioned earlier are the ratios of the recorded spectrum.
If a problem is well-posed it means among other things that the solution of the problem is unique. Some problems are however not well-posed, which is why many have looked into solving so called ill-posed problems Hansen (1998) where the covariate matrix does not have full rank. This will always be the case when there are more variables than observations ($p > n$). If such a problem is to be solved properly using Least Squares (LS), some sort of regularization is necessary. Typically this involves including additional assumptions, such as smoothness of the solution. Tikhonov regularization Tikhonov (1963) also known as ridge regression Hoerl and Kennard (1970) is one of the most common ways of regularizing a linear ill-posed problem or an overdetermined system. The ridge regression minimizes the residual sum of squares like an OLS, but in addition it penalizes the L2-norm of the model coefficients. This means all variables are kept in the model but in a smoothed manner. However, in some situations where $p >> n$, ridge regression is not well suited because it creates very complex and

thus very little interpretable models. This also means that if some variables contain none or little information regarding the dependent variable, they will still contribute to the final model and thus induce noise. Another approach to solve $p >> n$ problems is by using subset selection or stepwise selection. These methods choose variables having largest partial correlation with the dependent variable, and discards the rest. This type of model is also sometimes known as a parsimonious model and is often much more interpretable, although unfortunately often yields lesser prediction ability.
The Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Tibshirani in Tibshirani (1996) was created to solve this problem. Here an L1-norm penalization of the coefficients is used instead of the L2-norm. This means that a sparse solution, as is the case with stepwise selection, is obtained while still continuously smoothing the coefficients to some degree for good prediction. This approach proved to be an improvement of the ridge regression in many cases, while boosting regression and forward stagewise regression both were invented as alternative methods approximately thereafter. These are all described in Hastie et al. (2009).
A method able to obtain the solution of all these methods in a computationally fast manner is the Least Angle Regression (LARS)Efron et al. (2004), proposed by Efron. This regression method gives rise to at most the same amount of calculations as an ordinary LS. An alternative regularization and variable selection method is the elastic net (EN) by Zou Zou and Hastie (2005), which often outperforms forward stagewise regression as well as lasso regression. The elastic net can be incorporated into the LARS regression, commonly known as LARS-EN, and penalizes the L1 as well as the L2 norm of the coefficients; see (G.1).

$$\begin{aligned} L(x,\theta) &= \sum_{i=1}^{n}\left(\sum_{j=1}^{p}(\theta_j x_{ij}) - y_i\right)^2 \\ s.t. &\sum_{i=1}^{p}\theta^2 \leq s_1 \text{ and } \sum_{i=1}^{p}|\theta| \leq s_2 \end{aligned} \tag{G.1}$$

L denotes the loss function, which is the residual sum of squares. θ are the model coefficients and y is the dependent variable, in this case the experimental days. s_1 and s_2 are the constraint bounds on the LASSO and ridge constraints respectively, which together gives the Elastic Net constraint.
The Contours as well as constraints of a simple 2 dimensional problem, simulated as an example of (G.1) is also seen graphically in figure G.3. This regression scheme is especially suited to solve $p >> n$ problems due to the stability of the L2 norm, and the sparsity property of the L1 norm, which is also shown in Clemmensen et al. (2009b). Here the LARS-EN efficiently manages to select a set of suiting variables to detect water in different types of sand, which also is the reason why we have chosen to use it to solve the problem in this paper.

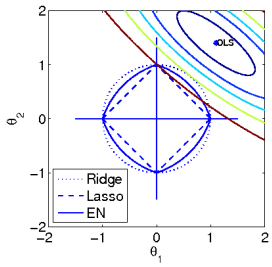


Figure G.3: Contours of the residual sum of squares function with the Ordinary LS solution defined as the minimum. The ridge, LASSO and Elastic Net constraints are similarly illustrated. Where the respective constraints and contours intersect, the ridge, LASSO and Elastic Net solutions will be defined.

G.4 Results and discussion

In order to generalize the model as much as possible a leave one out cross validation (LOO-CV)Hastie et al. (2009) was used on a training set (A) to estimate the model, and a separate test set (B) was used to evaluate the performance of the model. To further check the repeatability of the model, the training and test set were switched and a new model estimated. Predictions of the estimated models are seen in figure G.4. The boxes in the figures are standard type boxplots and have lines at the 25th, 50th and 75th percentiles. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. By visually inspecting the boxplots in figure G.4 there seems to be a tendency that the celeriac models having MSE: $(5.88, 4.77)$ respectively, generally have a better prediction ability than the carrot models having MSE: $(17.31, 13.88)$. It also seems that there is a difference between some of the groups in each of the four models, which generally increases slightly in the beginning and then flattens out towards the end. Specifically for the carrots it seems that after day 4, the predictions starts to oscillate, as if an equilibrium has been reached. The same is the case for the celeriac after day 8.
Figure G.5 shows the result of all pairwise two-sided bonferroni corrected t-tests between all days in each model. The bonferroni corrected t-tests test the H_0-hypothesis, that two groups can be assumed to come from the same population at the 5% level of significance. The statistical tests show that for carrots we are able to verify a significant change in the mean from day 2 to day 4. After day 4 we are not able to verify a significant change in mean for the carrots, which could indicate a steady state has been reached. However, for the celeriac we are able to significantly track a change from day to day until day 12, with the

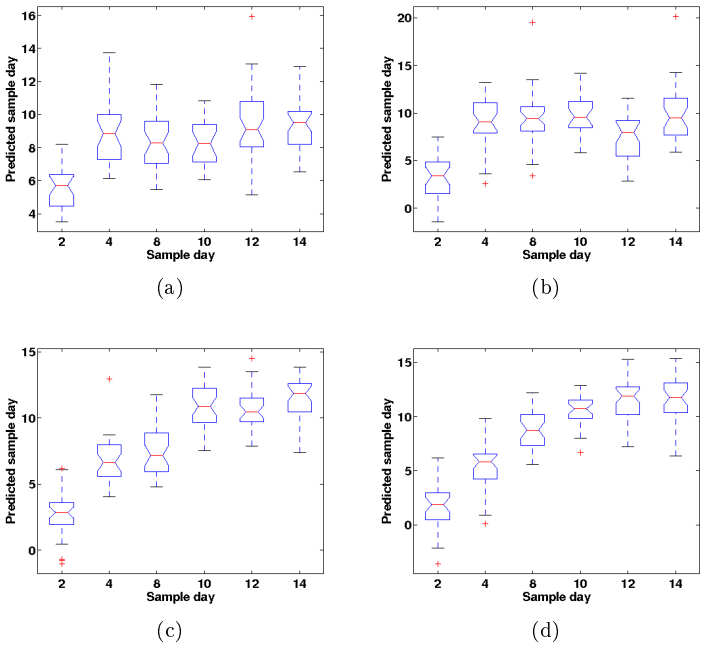


Figure G.4: All figures show predictions grouped by true sample day. The top two plots show predictions for the carrot samples, while the bottom two plots show the predictions of the celeriac. The two leftmost figures show the model trained with LOO-CV on dataset A and tested on dataset B, while the rightmost figures show the model trained with LOO-CV on dataset B and tested on dataset A.

	D2	D4	D8	D10	D12	D14
D2	0	2	2	2	2	2
D4	2	0	0	0	2	1
D8	2	0	0	0	2	0
D10	2	0	0	0	2	1
D12	2	2	2	2	0	2
D14	2	1	0	1	2	0

(a)

	D2	D4	D8	D10	D12	D14
D2	0	2	2	2	2	2
D4	2	0	1	2	2	2
D8	2	1	0	2	2	2
D10	2	2	2	0	1	1
D12	2	2	2	1	0	0
D14	2	2	2	1	0	0

(b)

Figure G.5: Both tables show all pairwise 2-sided bonferroni corrected t-tests between all groups of the carrot model(a) and the celeriac model(b). 0 indicates neither model can reject the H_0-hypothesis (i.e. we accept that the means are equal). 1 indicates one of the two models rejects the H_0-hypothesis, and 2 indicates both models reject the H_0-hypothesis(i.e. the groups are significantly different at a 5% level). Both tables are symmetrical respectively. The diagonal contain only zeros.

exception of day 8 where some uncertainty appears.
As mentioned in the section about the experimental setup, the vegetables were kept in the refrigerator in plastic bags. Plastic bags are not able to isolate oxygen molecules, which is why we believe the change in the spectra is caused by oxidation of the vegetables. An oxidation causes browning/graying of celeriac and carrots to become more pale. An increasing brown/gray color is a change in a wide range of the spectrum, and is essentially a change of brightness. The most significant components describing the celeriac consists of wavelengths from the entire visible spectrum, which coincides with a general shift in brightness. For the carrots the components seem to have a tendency to lie in the red/NIR area, which also coincides with a general more pale appearance, or removal of redness/orangeness, which essentially is an oxidation of the beta-carotene. This is exactly the color change to expect in an oxidation process of these vegetables.

G.5 Conclusions

An objective measure of the quality change of carrots and celeriac was proposed which uses multispectral image analysis. Six images were recorded over 14 days, for two different data sets. Each carrot or celeriac piece was isolated using a combination of a Canonical Discriminant Analysis and watershed algorithm, for a total of around 200 pieces per training and test set, for both carrots and celeriac pieces respectively. A set of 3249 features were extracted for each vegetable piece, giving rise to a very ill posed $p >> n$ problem. A special regression technique, Least Angle Regression-Elastic Net, was performed on both the carrot and celeriac data sets. Test and training were interchanged, resulting in two models per vegetable type. These were estimated to check the repeatability and statistical tests were performed to check if it in fact was possible to discriminate between the different days predicted on behalf of the estimated models.
The results showed that the celeriac predictions were somewhat better than the carrots, although a trend was seen in both. We see that there is a large change from day 2 to day 4 in the reflectance spectrum for both carrots and celeriac, and for the celeriac we see the change continuing until day 12. The pairwise two sided bonferroni corrected t-tests showed exactly that these changes were statistically significant at a 5% level of significance. The corresponding sensory tests showed no difference over the 14 days, which makes it the more important that we are able to detect minor changes using multispectral imaging.

G.6 Acknowledgements

The authors would like to thank Rene Thrane and Peter Reimer Stubbe for carrying out the manual experiments in the laboratory at the National Food Institute of Denmark. This study was financially supported by The Danish Food Industry Agency

Bibliography

Adler-Nissen, J., 2002. The continuous wok - a new unit operation in industrial food processes. Journal of Food Process Engineering 25, 435–453.

Adler-Nissen, J., 2005. Industrial stir frying. Asia Pacific Food Industry 17 (5), 32–34.

Adler-Nissen, J., 2007. Continuous wok-frying of vegetables: Process parameters influencing scale up and product quality. Journal of Food Engineering 83, 54–60.

Alon, A., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., , Levine, A. J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proceedings of the National Academy of Sciences, PNAS. Vol. 96. pp. 6745–6750.

Bao, H. N. D., Arason, S., Porarinsdottir, K. A., 2007. Effects of dry ice and superchilling on quality and shelf life of arctic charr (salvelinus alpinus) fillets. International Journal of Food Engineering 3 (3), article 7.

Bastien, P., Vinzi, V. E., Tenenhaus, M., 2005. Pls generalised linear regression. Computational Statistics and Data Analysis 48, 17–46.

Belhumeur, P., Hespanha, J., Kriegman, D., July 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence 19 (7), 711–720.

Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434.

Bellman, R., 1961. Adaptive control processes: A guided tour. Princeton university press.

Ben-Israel, A., Greville, T. N. E., 2003. Generalized Inverses, 2nd Edition. Springer.

Berge, A., 2007. Improving hyperspectral classification by simplified parameter estimates. Ph.D. thesis, University of Oslo, Department of Informatics.

Berman, M., Bischof, L., Huntington, J., March 1999. Algorithms and software for the automated identification of minerals using field spectra or hyperspectral imagery. In: Thirteenth International Conference on Applied Geologic Remote Sensing.

Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., Huntington, J. F., OCTOBER 2004. Ice: A statistical approach to identifying endmembers in hyperspectral images. IEEE transactions on geoscience and remote sensing 42 (10).

Bickel, P., Levina, E., 2004. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. Bernoulli 6, 989–1010.

Bishop, C. M., 2006. Pattern recognition and machine learning. Springer.

Blees, F. C., November 1989. Method and apparatus for preparing concrete mortar. US Patent No. 4,881,819.

Boser, B., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, 144–152.

Breiman, L., January 2001. Random forests. Statistics Department University of California Berkeley.

Bro, R., July 2009. Applications of tensor methods in life sciences data, talk at The Techincal University of Denmark at the European Workshop on Modern Massive Data Sets.
URL `http://mmds.imm.dtu.dk/`

Bu, H.-L., Li, G.-Z., Zeng, X.-Q., Yang, J., Yang, M., Oct. 2007. Feature selection and partial least squares based dimension reduction for tumor classification. In: IEEE International Conference on Bioinformatics and Bioengineering, BIBE. pp. 967–973.

Burgaard, M. G., Matzen, A., Adler-Nissen, J., 2004. Kontinuerlig wok til industriel brug. Plus Proces (6), 24–26.

Carver, R. L., August 1952. Method and apparatus for measuring moisture content of granular material. US Patent No. 2,607,830.

Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A., 2005. Discriminative common vectors for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1), 4–13.

Chan, K., Lee, T.-W., 2002. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. IEEE Transactions on Biomedical Engineering 49 (9), 963–974.

Chen, S., Donoho, D., Saunders, M., 1999. Atomic decomposition by basis pursuit. SIAM Journal of Scientific Computing 20 (1), 33–61.

Chun, H., Keles, S., 2009. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society - Series B.

Clemmensen, L., Hansen, M., Ersbøll, B., Frisvad, J., Jan 2007. A method for comparison of growth media in objective identification of penicillium based on multi-spectral imaging. Journal of Microbiological Methods 69, 249–255.

Clemmensen, L., Hastie, T., Ersbøll, B., 2009a. Sparse discriminant analysis. Technometrics(Resubmitted).

Clemmensen, L. H., Hansen, M. E., Ersbøll, B. K., 2009b. A comparison of dimension reduction methods with application to multi-spectral images of sand used in concrete. Machine Vision and Applications(Online first version).

Craw, I., Costen, N., Kato, T., Akamatsu, S., 1999. How should we represent faces for automatic recognition. IEEE Transactions on Pattern Analysis and Machince Intelligence 21 (8), 725–736.

Darkner, S., Paulsen, R. R., Larsen, R., Oct 2007. Analysis of deformation of the human ear and canal caused by mandibular movement. In: Proceedings of Medical Image Computing and Computer Assisted Intervention, MICCAI. Springer Lecture Notes.

Daugman, J., 1993. High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (11), 1148–1161.

Daugman, J., 2002. How iris recognition works. Vol. 1.

Ding, C., Li, T., Jordan, M., 2006. Convex and semi-nonnegative matrix factorizations. Tech. Rep. 60428, Lawrence Berkeley National Laboratory.

Donoho, D., 2006. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59 (6), 797–829.

Donoho, D. L., 2000. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, to the American Mathematical Society 'Math Challenges of the 21st Century'. Avialable from http://www-stat.stanford.edu/~donoho.

Drori, I., Donoho, D., 2006. Solution of l1 minimization problems by lars/homotopy methods. In: IEEE International Conference on Acoustics, Speech, and Signal Processing.

Dryden, I., Mardia, K., 1998. Statistical Shape Analysis. Wiley series in probability and statistics.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. John Wiley & Sons.

Efron, B., Hastie, T., Johnstore, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Elden, L., 2007. Matrix methods in data mining and pattern recognition. SIAM.

Elden, L., Park, H., 1999. A procrustes problem on the stiefel manifold. Numerische Mathematik 82, 599–619.

Ersbøll, B. K., Conradsen, K., 2003. An introduction to statistics. Vol. 2. IMM / Informatic and Mathematical Modelling.

Fergus, R., Weiss, Y., Torralba, A., 2009. Semi-supervised learning in gigantic image collections. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 522–530.

Figeuiredo, M. A. T., Jain, A. K., 2001. Bayesian learning of sparse classifiers. Vol. 1. p. 35.

Fisher, R., 1936. The use of multiple measurements in axonomic problems. Annals of Eugenics.

Fodor, I. K., Kamath, C., 2001. Dimension reduction techniques and the classification of bent double galaxies. Computational Statistics & Data Analysis 41, 91–122.

Fredriksson, T., Petersson, U., 1978. Severe psoriasis - oral therapy with a new retinoid. Dermatologica 157, 283–41.

Goldberg, D., 1989. Genetic algorithms in search, optimization and machine learning. Addison and Wesley.

Goldstein, A. J., Harmon, L. D., Lesk, A. B., 1971. Identification of human faces. In: Proceedings of the IEEE. Vol. 59. pp. 748–760.

Golland, P., 2001. Discriminative direction for kernel classifiers. In: Proceedings of Neural Information Processing Systems, NIPS. pp. 745–752.

Gomez, D. D., Clemmensen, L. H., Ersbøll, B. K., Carstensen, J. M., 2007. Precise acquisition and unsupervised segmentation of multi-spectral images. Computer Vision and Image Understanding 106 (2-3), 183–193.

Gonzalez, R. C., Woods, R. E., Eddins, S. L., 2002. Digital Image Processing, 2nd Edition. Prentice Hall.

Gonzalez-Salgado, A., Patino, B., Vazquez, C., Gonzalez-Jaen, M. T., 2005. Discrimination of aspergillus niger and other aspergillus species belonging to section nigri by pcr assays. FEMS Microbiology Letters 245, 353–361.

Grosenick, L., Greer, S., Knutson, B., December 2008. Interpretable classifiers for fmri improve prediction of purchases. IEEE transactions on neural systems and rehabilitation engineering 16 (6), 539–548.

Guigue, V., Rakotomamonjy, A., Canu, S., 2005. Kernel basis pursuit2. In: European Conference on Machine Learning, Porto.

Guo, Y., Hastie, T., Tibshirani, R., 2007. Regularized linear discriminant analysis and its applications in microarrays. Biostatistics 8 (1), 86–100.

Hand, D. J., 2006. Classifier technology and the illusion of progress. Statistical Science 21 (1), 1–15.

Hansen, P. C., 1998. Rank-Deficient and Discrete Ill-Posed Problems. SIAM.

Hastie, T., Buja, A., Tibshrani, R., 1995a. Penalized discriminant analysis. The Annals of Statistics 23 (1), 73–102.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd Edition. Springer.

Hastie, T., Tibshrani, R., 1996. Discriminant analysis by gaussian mixtures. Journal of Royal Statistical Society - Series B 58, 158–176.

Hastie, T., Tibshrani, R., Buja, A., 1995b. Flexible discriminant and mixture models. In: Neural Networks and Statistics conference, Edinburgh. J. Kay and D. Titterington, Eds. Oxford University Press.

Hawksworth, D. L., 2001. The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycological Research 105, 11422–1432.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gutsterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O. P., Borg, A., Trent, J., 2001. Gene-expression profiles in hereditary breast cancer. The New England Journal of Medicine 344, 539–548.

Hesterberg, T., Choi, N. H., Meier, L., Fraley, C., 2008. Least angle and l1 penalized regression. Statistics Surveys 2, 61–93.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Jolliffe, I. T., 1986. Principal Component Analysis. Springer, New York.

Juhasz, A., Pfeiffer, I., Keszthelyi, A., Kucsera, J., Vagvolgyi, C., Hamari, Z., 2008. Comparative analysis of the complete mitochondrial genomes of aspergillus niger mtdnatype1a and aspergillus tubingensis mtdnatype 2b. FEMS Microbiology Letters 281, 51–57.

Karush, W., 1939. Minima of functions of several variables with inequalities as side constraints. Master's thesis, University of Chicago, Department of Mathematics.

König, A., 2000. Dimensionality reduction techniques for multivariate data classification, interactive visualization, and analysis - systematic feature selection vs. extraction. In: Proceedings of Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Vol. 1. IEEE, pp. 44–55.

Larsen, R., Olafsdottir, H., Ersbøll, B., 2009. Shape and texture based classification of fish species. In: 16th Scandinavian conference on image analysis. Springer Lecture Notes in Computer Science.

Leardi, R., 2000. Application of genetic algorithm-pls for feature selection in spectral data sets. Journal of Chemometrics 14, 643–655.

Leardi, R., Gonzalez, A. L., 1998. Genetic algorithms applied to feature selction in pls regression: how and when to use them. Chemometrics and intelligent laboratory systems 41, 195–207.

Lee, C. M., Narayanan, S., Pieraccini, E., 2001. Recognition of negative emotions from the speech signal. In: 2001 IEEE Workshop on Automatic Speech Recognition and Understanding. ASRU 2001. Conference Proceedings. pp. 240–243.

Lee, D., Seung, H., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–91.

Lee, D., Seung, H., 2000. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems, 556–562.

Leistner, C., Saffari, A., Santner, J., Bischof, H., 2009. Semi-supervised random forests. In: Proceedings of International Conference on Computer Vision, ICCV.

Leng, C., 2008. Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data. Computational biology and chemistry 32, 417–425.

Li, F., Yang, Y., Xing, E., 2006. From lasso regression to feature vector machine. In: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA, pp. 779–786.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. Kluwer Academic Publishers.

M. Aizerman, E. B., Rozonoer, L., 1964. Automation and Remote Control 25, 821–837.

Messer, K., Kittler, J. M. J., Luettin, J., Maitre, G., 1999. Xm2vtsbd: The extended m2vts database. In: Proceedings of 2nd Conference on Audio and Video-base Biometric Personal Verification, AVBPA. Springer Verlag, New York.
URL `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb`

Mørup, M., Clemmensen, L., 2007. Mulasso.
URL `http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/5235/zip/imm5235.zip`

Osborne, M., Presnell, B., Turlach, B., 2000. A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis 20 (3), 389–403.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9 (1), 62–66.

Petersen, L., Minkkinen, P., Esbensen, K. H., 2005. Representative sampling for reliable data analysis: Theory of sampling. Chemometrics and Intelligent Laboratory Systems 77, 261–277.

Philip, P. J., Moon, H., Rizvi, S. A., Rauss, P. J., 2000. The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10), 1090–1104.

Pochet, N., Smet, F. D., Suykens, A. K., , Bart, L. R. D. M., 2004. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics 20 (17), 3185–95.

Rasmussen, G., 2006. Hr-mas nmr data acquisition and chemometric analysis of fungal extracts. Master's thesis, E06, BioCentrum, Technical University of Denmark.

Rawlings, J. O., 1998. Applied regression analysis - a research tool. Wadsworth & Brooks/Cole.

Razek, J., August 1989. Method and apparatus for measuring moisture content of sand or the like. US Patent No. 4,853,614.

Rencher, A. C., 2002. Methods of Multivariate Analysis. John Wiley & Sons.

Schuster, E., Dunn-Coleman, N., Frisvad, J. C., van Dijck, P. W. M., 2002. On the safety of aspergillus niger - a review. Applied Microbiology and Biotechnology 59, 426–435.

Sha, F., Saul, L., Lee, D., 2002. Multiplicative updates for nonnegative quadratic programming in support vector machines. In: Advances in Neural Information Processing Systems 15.

Shaobing, S., Donoho, D., 1994. Basis pursuit. In: Proceedings of 28th Asilomar conference on Signals, Systems and Computers.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, UK.

Shi, J., Samal, A., Marx, D., 2006. How effective are landmarks and their geometry for face recognition? Computer Vision and Image Understanding 102, 117–133.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, H., Redwine, E., Yang, N., 1989. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii. radical prostatectomy treated patients. Journal of Urology 16, 1076–1083.

Szepietowski, J. C., Sikora, M., Pacholek, T., Dmochowska, A., 2001. Clinical evaluation of the self-administered psoriasis area and severity index (sapasi). dermatovenerologica - alpina, pannonica et adriatica 10 (3), 1–7.

Thodberg, H. H., Ólafsdóttir, H., sep 2003. Adding curvature to minimum description length shape models. In: British Machine Vision Conference, BMVC.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society - Series B 58 (No. 1), 267–288.

Tibshirani, R., Saunders, M., 2005. Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society - Series B 67 (1), 91–108.

Tikhonov, A. N., 1963. Solutions of incorrectly formulated problems and the regularization method. Soviet Mathematics Doklady 4, 1035–1038, english translation of Dokl. Akad. Nauk. SSSR, 151:501-504, 1963.

Trendafilov, N. T., Jolliffe, I. T., 2007. Dalass: Variable selection in discriminant analysis via the lasso. Computational statistics and data analysis 51, 3718–3736.

Turk, M., Pentland, A., 1991. Face recognition using eigenfaces. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Vapnik, V., 1999. The Nature of Statistical Learning Theory, 2nd Edition. Springer-Verlag, New York.

Vina, S. Z., Chaves, A. R., 2003. Texture changes in fresh cut celery during refrigerated storage. Journal of the Science of Food and Agriculture 83, 1308–1314.

Viola, P., Jones, M., 2001. Robust real-time object detection. In: Proceedings of IEEE Workshop on Statistical and Computational Theories of Vision.

Walker, J., 1982. Uses of neutrons in egineering and technology. Physics in Technology 13, 239–48.

Wang, L., Zhu, J., Zou, H., 2006. The doubly regularized support vector machine. Statistica Sinica 16, 589–615.

Wu, D., Boyer, K. L., 2009. Resilient subclass discrimiannt analysis. In: Proceedings of International Conference on Computer Vision, ICCV.

Ye, J., 2007. Least squares linear discriminant analysis. In: Proceedings of the 24th International Conference on Machine Learning, ICML. pp. 1087 – 1093.

Yeoh, E.-J., , et. al, March 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143.

Zhang, S., Wang, W., Ford, J., Makedon, F., 2006. Learning from incomplete ratings using non-negative matrix factorization. In: Proceedings of 6th SIAM Conference on Data Mining, SDM.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society - Series B 67 (Part 2), 301–320.

Zou, H., Hastie, T., Tibshirani, R., June 2006. Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286.

Index